Multimed Tools Appl (2013) 62:51–73
DOI 10.1007/s11042-011-0916-9

Classifying cinematographic shot types

Luca Canini · Sergio Benini · Riccardo Leonardi

Published online: 3 November 2011

Abstract In film-making, the distance from the camera to the subject greatly affects the narrative power of a shot. By the alternate use of Long shots, Medium and Close-ups the director is able to provide emphasis on key passages of the filmed scene. In this work we investigate five different inherent characteristics of single shots which contain indirect information about camera distance, without the need to recover the 3D structure of the scene. Specifically, 2D scene geometric composition, frame colour intensity properties, motion distribution, spectral amplitude and shot content are considered for classifying shots into three main categories. In the experimental phase, we demonstrate the validity of the framework and effectiveness of the proposed descriptors by classifying a significant dataset of movie shots using C4.5 Decision Trees and Support Vector Machines. After comparing the performance of the statistical classifiers using the combined descriptor set, we test the ability of each single feature in distinguishing shot types.

Keywords Shot type · Movie content · Feature extraction

1 Introduction

When watching movies, the feeling is that some film directors have sharply different styles that are easily recognisable. These individual styles can be identified not only

L. Canini (✉) · S. Benini · R. Leonardi
Department of Information Engineering, University of Brescia, Brescia, Italy
e-mail: luca.canini@ing.unibs.it

S. Benini
e-mail: sergio.benini@ing.unibs.it

R. Leonardi
e-mail: riccardo.leonardi@ing.unibs.it

in the content, but also from the formal aspects of the films. In cinematography in fact, a widely accepted set of directing rules are often adopted to link the meanings of the film shot to be conveyed with various camera-related attributes.

As proposed in [33], the obvious approach in searching for individual characteristics in the formal side of a director's grammar is to consider those variables that are most directly under the director's control. These are also, to a certain extent, those that are the easiest to quantify, such as *shot length*, meant as shot duration, *shot type* in terms of closeness of the camera to the subject, *camera movement* such as pan, tilt, zooms, *shot transitions* (cut, fades, dissolves, wipes), etc.

While a certain amount of work has been done in investigating most of these characteristics (as in the exhaustive study in [38]), so far not much attention has been specifically directed towards automatic identification of the shot type, that is related to the distance between camera and the main recorded subject [1].

Varying the camera distance from the subject of interest is a common directing rule used to subtly adjust the relative emphasis between the filmed subject and the surrounding scene [38]. Although the gradation of distances is infinite, in practical cases the categories of definable shot types can be re-conducted to three fundamental ones: *Long shots* (LS), *Medium shots* (MS), and *Close-ups* (CU).

A Close-up shows a fairly small part of the scene, such as a character's face, in such a detail that it almost fills the screen. This shot abstracts the subject from a context, focusing attention on a person's feelings or reactions, or on important details of the story. Different grades of Close-up are presented in Fig. 1a, depicting human characters from the breast upwards.

In a Medium shot, as in the case of the standing actors depicted in the examples of Fig. 1b, the lower frame line passes through the body from the waist down to include the whole body (in this case it is called *Full shot*). In such shots, the actor and the setting occupy roughly equal areas in the frame, while leaving space for hand

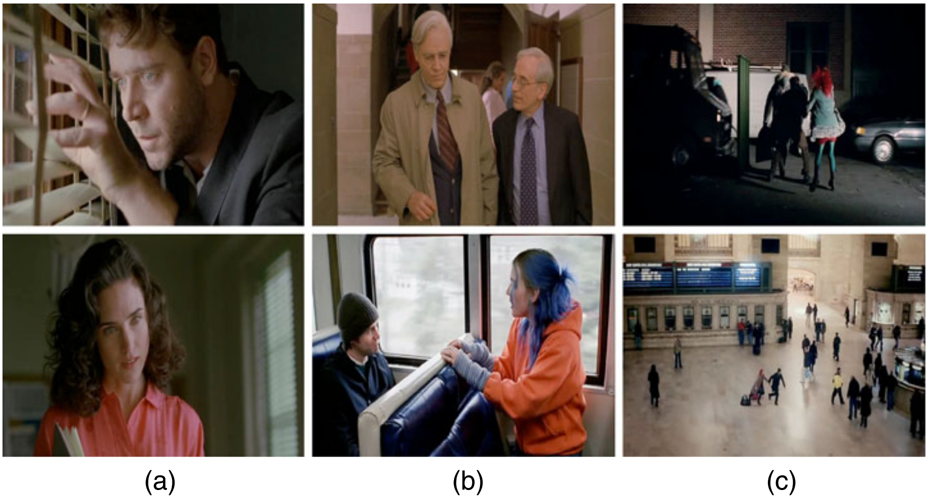


Fig. 1 Shot types: **a** Close-ups, **b** Medium and **c** Long shots, as in [1]. Images are from "A Beautiful Mind" and "Eternal Sunshine of the Spotless Mind"

gestures to be seen. Medium shots are also frequently used for the tight presentation of two actors, or with dexterity, three.

Finally, Long shots show all or most of a fairly large subject (for example, a person) and usually much of the surroundings. This category comprises also Extreme Long shots (as shown in Fig. 1c) where the camera is at its furthest distance from the subject, emphasising the background, often used as the opening shot of a sequence to set the scene (also called *Establishing shot*). The reader can refer to [1] for a more detailed taxonomy on shot types.

Of course camera distance is just part of a greater taxonomy of movie stylistic capabilities; these include, among the others, visual features (such as *colour*, *framing*, *lightning*, *composition*) as well as sound and higher-order entities (*e.g*, *rhythm*, *editing*, *continuity/discontinuity*), etc. For a more complete view on the topic and to better establish the context of the study, the interested reader can refer to [1, 5], or [25].

1.1 Paper aims and organisation

In this paper we investigate five techniques which study intrinsic characteristics and content of single shots containing indirect information about camera distance from the focus of attention, and we use them for classifying shots into the three categories (LS, MS or CU).

The first technique investigates the colour intensity distribution on local regions in frames. A second technique employs *Motion activity maps* [40] which, computing the accumulation measurement of motion activity on the grids of shot frames along the time axis, estimate the occupancy of the space by moving foreground objects. The third method relies on the geometry of the scene, by measuring the angular aperture of perspective lines found by Hough transform. The fourth measure relies on actual shot content, by detecting faces in frames: face dimensions, estimated by a well-know detection algorithm [37], provide an indirect measure of the absolute distance between the camera and the filmed subject. Finally, by inspecting in the frequency domain the spectral amplitude of the scene and its decay, it is possible to discriminate between different image structures and their spatial scales.

These methods take into consideration only one aspect at a time of the shot, i.e., its *colour intensity properties*, its *motion distribution*, its *geometry*, its *content* and its *spectral component*, so when considered singularly, they may not be accurate enough for a robust classification into shot types. For this reason, the combined set of descriptors is adopted to feed two supervised statistical classifiers, namely C4.5 decision trees [30], and Support Vector Machine (SVM) [12]. After comparing advantages and drawbacks of the two classification approaches, the ability of single descriptors in categorising shot types is also explored.

The main advantage of the described approach lies in the fact that the proposed method works on frames directly extracted from the filmed video sequence, by combining multiple easy-to-obtain features for fast and robust classification. Differently from other techniques, there is no need to compare different images of the same scene to draw spatial information. Furthermore, the proposed scheme can be applied to narrative video genres (e.g. films), which show high variability in the showed content, and its validity is not limited as most of the prior work, to the analysis in the sport domain, which allows for easier application of colour cues to recover camera distance.

The performed analysis could be beneficial to applications of semantic content analysis and editing, video retrieval, summarisation and, as emerged in the last few years, of affective analysis of feature films [16]. In fact, it is often through different combinations of shot properties that a director defines his/her style, as well as captivate and drive the attention of the viewers, allowing the film's intentions to be properly conveyed [38]. For example, a possible application based on this framework can be envisaged for studying the relationships between the usage of different patterns of shot types in movies and the affective reaction of a large community of viewers.

This document is organised as follows. In Section 2, the existing literature on the topic is reviewed. In Sections 3–7 the five aforementioned features containing indirect information about camera distance from the focus of attention are described. Section 8 first discusses the composition of the database from which these characteristics are extracted; then the adopted classification approaches (SVM and C4.5 decision tree) are described, tested, and results discussed. Considerations on future work and conclusions are finally drawn in Section 9.

2 Previous work

Techniques able to directly estimate the shot type starting from single images are not numerous. Not surprisingly, a number of them focuses on the automatic classification of shot types coming from sport videos, where the type of the filmed shot is often less relevant than in feature movies, at least from the narrative perspective. In soccer videos, as analysed in [39], the difference between shot types is useful for distinguishing *plays* from *breaks*, and it is determined investigating the ratio of green grass area in shot frames. By using dominant colour ratio as an effective feature, authors distinguish Long shots, which have the largest grass area, Medium ones, which have less, and Close-ups which have hardly any. Similar approaches based on grass presence and domain modelling are presented in other works on sport videos, such as in [13] and in [15].

An alternative approach to infer the shot type could rely on measurements of scene depth, specifically by estimating the distance between the camera and the main filmed subject. Literature on *absolute* depth estimation (i.e., the actual distance between the camera and the subject) is very large, but the proposed methods rely on a limited number of sources of information (e.g., binocular vision, motion parallax, or defocus). As pointed out in [36], when looking at a photograph, human observers can provide a rough estimate of the absolute depth of a scene even in the absence of all these sources of information. Therefore, in the same work [36], the authors estimate the absolute scene depth by recognising local and global spectral features of the structures present in the image. One alternative source of information for estimating the absolute depth of a shot is the size of recognisable objects contained in a scene, like faces, hands, cars, etc. as in [26]. Unfortunately, the required process of image segmentation and object recognition is often too computationally expensive and the outcoming classification remains still unreliable.

In general, when cues of absolute depth are absent, the distance between the observer and a scene cannot be estimated with a high degree of precision. However, the cinematographic denominations used for shot types (LS, MS, CU) do not necessarily imply an absolute distance [1]. This terminology deals with concepts, and it is obvious that the distance between camera and subject is different in a close shot of a house and in a close shot of a man.

For determining the shot type then, it could be also helpful to estimate the *relative* depth between scene elements. Currently available techniques able to estimate *relative* scene depth, on which to infer the shot type, mainly focus on shape from shading [7], texture gradients [35], edges and junctions [2], symmetrical patterns [34], fractal dimensions [21], and other pictorial cues such as occlusions, relative size, and elevation with respect to the horizon line [28]. The interpretation of shadows, edges and lines can be used for the reconstruction of a 3D model of the scene as in [17], but when taken alone, they difficultly bring some information about the scale of the scene itself.

To the best of our knowledge, before our initial analysis in [3] only other two works in literature directly dealt with shot type detection in movies. The work in [11] defines human body-based rules to extract the shot type from a limited set of 66 shots excerpted from movies. This system adopts a number of thresholds to filter the dimension and position of faces in video frames. As a consequence no decision can be taken when no actors are screened.

The work in [38] instead, proposes a systematic approach based on motion descriptors to build taxonomy for film directing semantics, where the camera distance from the focus of attention is used as an intermediate feature to distinguish *contextual-tracking* and *focus-tracking* shots. Even though the employed data corpus is in this case significant, the adopted classifier is binary, and uses Close-up-Medium and Long shots as classes, thus without distinguishing between CU and MS. Furthermore no classification performance is reported for this intermediate step of the work.

3 Local distribution of colour intensity

The first descriptor we propose aims at measuring the total percentage of pixels designated as background with respect to the frame area. Even if it is certainly true that the amount of background area is not strictly proportional to the camera distance, this descriptor based on local colour intensity histogram on images, allows for a coarse differentiation between camera distance categories.

The descriptor is computed on single key-frames extracted from the movie shots (any existing technique for shot boundary detection and key-frame extraction can be employed, without loss of generality) and it is based on the following considerations.

When looking to a picture, as pointed out in [10], it is quite easy to observe, for example in images representing landscapes, that edges of distant elements (such as mountains) are not as sharp as those of foreground objects. Due to the diffusion of rays of light in an opaque medium (such as air, which contains a great number of water particles, responsible for light diffraction), colours of distant elements tend to blend and generate a sort of blur. As a result, images become more and more uniform as the distance from the camera increases, and background colour appears as a weighted average of the colours present in the scene: in gray-level images, the zones which are perceived as more blurred (i.e., which are farther from the camera) have gray levels gathered around an average value. On the contrary, in those areas where edges are sharper (i.e., nearer to the camera) gray levels are more scattered.

The algorithm here proposed confirms these intuitions exposed in [10] and develops a more complete criterion for camera distance estimation. The analysis is performed on the basis of the second order statistics of local image histograms. First each colour key-frame of dimension $\mathcal{W} \times \mathcal{H}$ is converted into the corresponding one-channel gray-level image $I(x, y)$. A local histogram is then computed over a rectangular sliding window w_I of dimensions $\frac{\mathcal{W}}{R} \times \frac{\mathcal{H}}{R}$ (where $\mathcal{R} = 20$, but is tuneable) centered on pixel $(\overline{x}, \overline{y})$ and scanning $I(x, y)$. Indicating with $f(g, w_I)$ the number of pixels in the window w_I whose gray level is equal to g, the average gray level of the histogram computed on the window w_I is obtained as:

$$\overline{g}_{w_I(\overline{x},\overline{y})} = \frac{\sum_i g_i \cdot f(g_i, w_I)}{\sum_i f(g_i, w_I)}$$

and its variance $\sigma^2_{w_I(\overline{x},\overline{y})}$ is computed as:

$$\sigma^2_{w_I(\overline{x},\overline{y})} = \frac{\sum_i (g_i - \overline{g}_{w_I(\overline{x},\overline{y})})^2 \cdot f(g_i, w_I)}{\sum_i f(g_i, w_I)}$$

On this basis, the *histogram variance* image $I_\sigma(x, y)$ is created. This is a gray-level image, where the value of pixel $(\overline{x}, \overline{y})$ is given by the variance $\sigma^2_{w_I(\overline{x},\overline{y})}$ of the histogram computed on the window scanning $I(x, y)$ and centered in $(\overline{x}, \overline{y})$. Variance values are then normalised to the maximum obtained on an entire set of movie key-frames in the range $[0, 255]$. In the obtained image $I_\sigma(x, y)$, scene points likely farther with respect to the camera will be the darkest ones (those with lowest variance), while

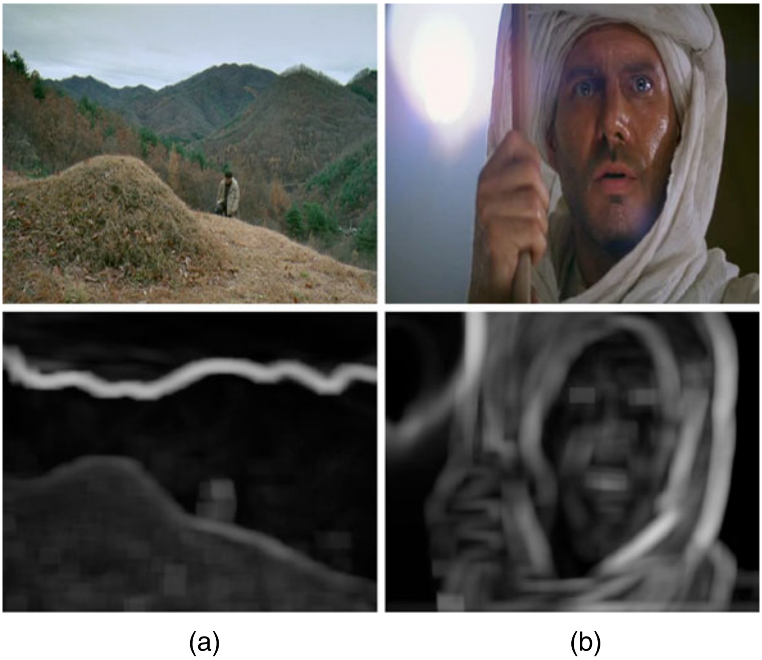


Fig. 2 Examples of original images and the obtained histogram variance images for **a** a Long shot from “Samaritan Girl” and **b** a Close-up from “Raiders of the Lost Ark”

image zones supposedly closer to the observer will be brighter (those with higher variance). Examples of original images and the obtained histogram variance images are shown in Fig. 2 for a Long shot in (a) and a Close-up in (b).

It is certainly true that not all high variance pixels always belong to the foreground area, e.g., the high intensity line dividing the sky from the mountains in Fig. 2a. However the obtained image functions adequately as a detector of background areas useful for camera distance categorisation.

The scalar value of the descriptor is finally obtained by a two-step process. First a binary segmentation on $I_\sigma(x, y)$ assigns black value to "farther" pixels and white value to "closer" ones, where the threshold used for the segmentation is adaptively set according to the method exposed in [23]. Then, after excluding too small connected components under a minimum area, the ratio A_σ of all black connected components to the total frame area is computed. A_σ provides an indirect estimation of camera distance, since it estimates the amount of background present in the key-frame, which is useful for the classification of the shot type: Long shots have the largest background area, Medium ones have less, while Close-ups have hardly any.

4 Motion activity maps

Another criterion for camera distance estimation is derived from a motion descriptor able to characterise the perceived activity of motion in a shot, as well as its unique spatial distribution. Moving objects in the foreground are responsible for high *motion activity* (which describes the spatial distribution of the motion field modules [20]), since they occupy a large portion of the frame. On the contrary, a moving object pictured in a Long shot, due to its relative small dimension, do not contribute to a dramatic increase of motion activity.

At times when we are concerned with global motion and its spatial distribution in the scene, we can analyse motion of a video segment from the image plane along its temporal axis and generate the *Motion activity maps* (Mam) as in [40]. Used in the past for video indexing [4], Mam extracted from predicted frames of the MPEG-4 compressed stream are here adopted as an alternative source of information for estimating the occupancy of the frame space by moving foreground objects, thus providing an indirect measure of camera distance.

From each video shot $\mathcal{S}$ of frame dimension $\mathcal{W} \times \mathcal{H}$ a corresponding Mam image I_M with same dimensions is extracted. Each Mam is made up of $\frac{\mathcal{W}}{\mathcal{Q}} \times \frac{\mathcal{H}}{\mathcal{Q}}$ macroblocks of $\mathcal{Q} \times \mathcal{Q}$ identical pixels, where the value of $\mathcal{Q}$ depends on the adopted codec—typical values are $\mathcal{Q} = 4, 8, 16$. The value of each pixel (x, y) of I_M is the normalised numeric integral, computed over all predicted frames f_p of the shot $\mathcal{S}$, of the magnitudes of motion vectors $\overline{m}_v(\mathcal{B}_{i,j})$ associated to the macroblock $\mathcal{B}_{i,j}$ containing pixel (x, y), that is:

$$I_M(x, y) = \frac{1}{\# f_p} \sum_{f_p \in \mathcal{S}} \left| \overline{m}_v \left(\mathcal{B}_{i,j} \right) \right|_{f_p} \quad \text{s.t.} \ (x, y) \in \mathcal{B}_{i,j}$$

Therefore in a Mam, single pixel intensities measure the amount of motion undergone by the corresponding $\mathcal{Q} \times \mathcal{Q}$ macroblock $\mathcal{B}_{i,j}$ averaged over the shot duration, and normalised to a 8-bit representation over the entire set of movie shots. Such as other motion descriptors (e.g. MPEG-7 motion activity descriptor) this descriptor considers the overall intensity of motion activity in the scene, without distinguishing between the camera motion and the motion of the objects present in the scene.

As an example, in Fig. 3a a key-frame extracted from the movie "Raiders of the Lost Ark" is given, together with a representation of the associated motion field. In Fig. 3b, instead, the Mam extracted from the same shot is provided, where brightest regions correspond to high motion zones and darker ones are those which remain still during the shot.

The utility of a motion activity map is twofold: on the one hand, it indicates if the activity is spread across many regions or restricted to a large one, providing a view of the spatial distribution of motion. To a certain extent, motion activity maps thus admit an indirect measure of the number of moving objects, and hence the possibility to infer shot distance.

On the other hand, a Mam expresses the behaviour of motion activity in the shot by displaying its average temporal distribution over the shot duration. Combining these information we derive a clue about presence and rough dimensions of moving foreground objects. A binary segmentation process on $I_M(x, y)$ assigns black value to "still" macroblocks and white value to "active" ones. The indirect estimation of the camera distance is found by measuring the ratio A_M of all white connected components to the total frame area. The threshold used for the segmentation is adaptively set as in Section 3 and, to exclude too small connected components, only those with a minimum area are taken into account.

The computed descriptor A_M provides a cue on the amount of moving foreground objects in the shot and can be considered the dual descriptor with respect to the local distribution of colour intensity, which measured the amount of background occupation. Again, it is clear that the percentage of foreground moving objects is not strictly inversely proportional to camera distance. However it is still an adequate descriptor for a rough classification of the shot type: Long shots have the smallest foreground areas, Medium ones have bigger ones, while Close-ups mostly have foreground zones.

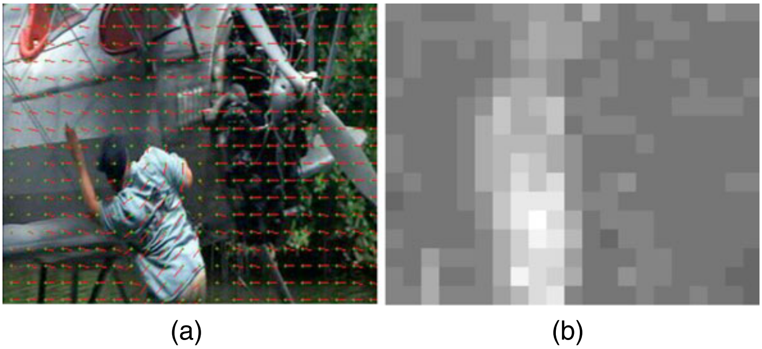


Fig. 3 **a** The motion vector field of a key-frame from "Raiders of the Lost Ark" and **b** its corresponding Mam

5 Scene perspective

This descriptor exploits the geometry of the scene to derive information about camera distance. In particular we are interested in detecting perspective lines in shot key-frames in order to estimate the distance from the focus of attention.

Hough transform [14] allows to detect segments, curves and predefined shapes in an image. The basic theory of the Hough line transform is that any point in a binary image could be part of a straight line. To check this, candidate line points are first extracted by an edge detector with Canny operator performed on the one-channel version of the image. Then, according to a probabilistic Hough transform, each point in the binary image is mapped into a locus of points in the Hough-plane, corresponding to all possible lines passing through that point. Summing over all contributions, lines that appear in the input image are local maxima in the Hough-plane (called the *accumulator plane*).

Any perspective representation of a scene that includes perpendicular lines has one or more vanishing points. Hough has been used already in [24] for detecting vanishing point in images. However the task is quite challenging, due to the variety of existing scenes. Perspectives consisting of many parallel lines are observed most often when shooting architectures or *man-made* environments (in this case it is not rare to see perspectives with several vanishing points). In contrast, *natural* scenes often do not have any sets of parallel lines and such a perspective would thus have no vanishing points.

Instead of tackling a precise detection of vanishing points, we rather aim at finding an estimator of camera distance by collecting slopes of perspective lines in shot key-frames. Since all lines parallel with the viewer's line of sight recede towards the vanishing point, perspective lines in a long shot remain parallel (due to the high distance from the vanishing point, as in Fig. 4a). Conversely, inclinations of perspective lines evidently differ when observing a less distant shot (Fig. 4b).

By measuring the angles at which perspective lines are inclined to the vertical axis $\overline{u}_y$ we are able to derive an estimator of the camera distance. Indicating with θ_i the angles of the n perspective lines whose angles with vertical are in the interval $0 < \theta_i < \pi/2$, and with ϕ_i the angles of the m lines whose angles with the vertical are in the interval $\pi/2 < \phi_i < \pi$, the average inclinations $\overline{\theta}$ and $\overline{\phi}$ are:

$$\overline{\theta} = \frac{1}{n}\sum_i \theta_i \qquad \text{and} \qquad \overline{\phi} = \frac{1}{m}\sum_i \phi_i$$

where we have ignored the vertical and horizontal lines, because of their non informativeness in terms of scene perspective. The *angular aperture* α of the perspective lines (for analogy with the angular aperture of lenses) is then given by the difference between the two average inclinations, that is:

$$\alpha = \frac{\overline{\phi} - \overline{\theta}}{2}$$

which provides a rough indicator of the distance between camera and the filmed subject.

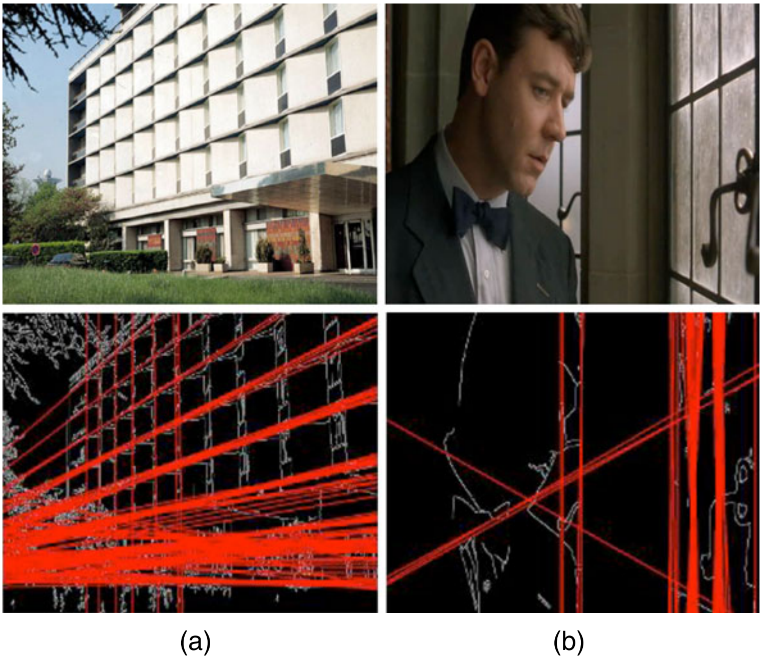


Fig. 4 Examples of perspective lines extracted by the Hough transform **a** from a Long shot and **b** from a Close-up taken from "A Beautiful Mind"

6 Faces and camera distance

A cue of absolute distance is provided when the size of a recognisable image object, e.g., a human face, is measurable.

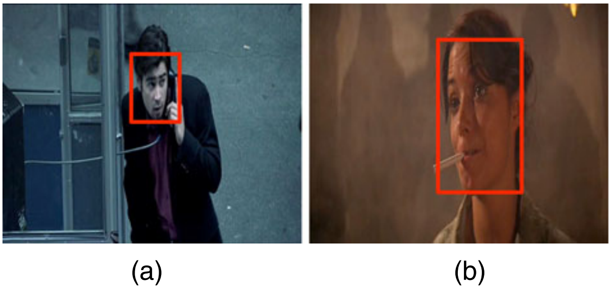
Apart from few (almost) people-less movies found in those filmic productions characterised by an abstract treatment of the space, such as in the early productions by Antonioni or Tarkovsky, the presence of human figures is central to modern cinematography, so that the probability of having a face in a scene is relevant.

While for other image objects the prior process of segmentation and recognition is still too computationally expensive, fast and robust detection algorithms for face detection do exist, such as the one in [37]. Despite the fact that this algorithm is known to work well for frontal faces only, the last implementation in [6] also comes with several cascade files for detecting profile faces, even if with slightly lower performance. In Fig. 5 an example of the output provided by the Viola-Jones method is given, where detected faces are highlighted by bounding boxes.

Since the descriptor here proposed is based on human body information, only shots containing actors are considered as relevant.

The descriptor A_F is computed as the ratio of the area occupied by the biggest bounding box to the total frame area. It provides an indirect measure of the shot type: Long shots have small bounding boxes, Medium ones have bigger ones, while Close-ups have a large portion of the frame covered by the detected face. In the specific example of Fig. 5, (a) is classified as MS, while (b) is a CU.

Fig. 5 Examples of faces (red bounding boxes) **a** in a MS from “Phone Booth” and **b** in a CU from “Raiders of the Lost Ark”



7 Global spectral amplitude

While previous techniques investigate intrinsic characteristics and the shot content in the pixel domain, we propose to complete the set of features looking at frame properties in the transform domain. As already suggested by [36], the magnitude of the global Discrete Fourier Transform (DFT) of an image $I(x, y)$ of dimension $\mathcal{W} \times \mathcal{H}$, defined as:

$$\left|\mathcal{I}\left(\overline{f}\right)\right| = \left|\sum_{x=0}^{\mathcal{W}-1}\sum_{y=0}^{\mathcal{H}-1} I(x, y)e^{-j2\pi\left(\frac{xf_x}{\mathcal{W}}+\frac{yf_y}{\mathcal{H}}\right)}\right|$$

contains information about the dominant orientations and spatial scale of the image.

Concerning real-world scenes, the shape of the spectral amplitude of the DFT is also very effective in revealing the spatial structure of the scene, allowing a clear distinction between *natural* versus *man-made* scenes, i.e. between images depicting natural subjects versus pictures mainly containing buildings, structures or objects built by humans. As shown in Fig. 6, while a natural scene presents an energy spectrum which is quite homogenous in all orientations with slight biases towards the horizontal and vertical orientations (Fig. 6a), a man-made one has sharp dominant vertical and horizontal components due to the presence of geometrical artificial structures (Fig. 6b).

The distinction between man-made and natural is also useful to study the scene scale, due to the fact that the spectral properties of these two kinds of images strongly differ as long as the distance of the camera increases [36].

To investigate the relationships between image structures and the distance D between the camera and the scene, it is useful to model the spectral amplitude of the DFT of (1) as proposed in [27]:

$$\left|\mathcal{I}\left(\overline{f}\right)\right| \sim \Lambda(D, \theta)/\left\|\overline{f}\right\|^{\Gamma(D,\theta)}$$

where θ is the phase of the frequency vector $\overline{f}$, $\Lambda(D, \theta)$ is a magnitude factor, and $\Gamma(D, \theta)$ is the slope which describes the decay of the spectral amplitude in logarithmic units.

By measuring the slopes of the spectral amplitude along the three main directions (vertical slope γ^v, horizontal slope γ^h and diagonal slope γ^d), we derive two vectors, one for natural images and one for man-made ones, respectively:

$$\Gamma_n = [\gamma_n^v \;\; \gamma_n^h \;\; \gamma_n^d] \quad \text{and} \quad \Gamma_m = [\gamma_m^v \;\; \gamma_m^h \;\; \gamma_m^d]$$

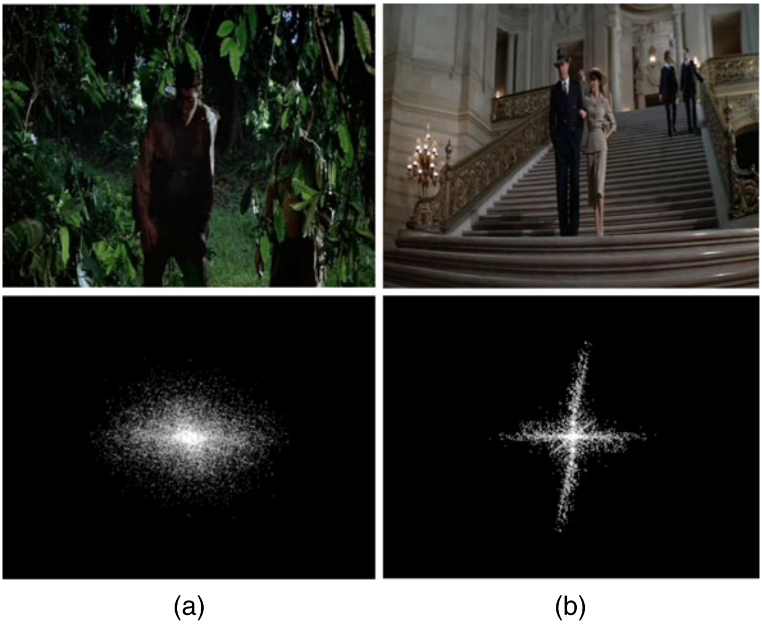


Fig. 6 Examples of global magnitude of the Fourier transform of **a** a natural and **b** a man-made image both taken from "Raiders of the Lost Ark" (the white plots represent the 80% of the energy)

which are helpful in estimating distance D and, more interesting for us, three families of values for D: LS, MS and CU.

Two different estimators are needed since as pointed out before, these two classes of images present different spectral and structural properties. In natural images, due to their irregular structure, the roughness of the picture diminishes on average with the distance, concentrating more energy in the lower frequencies. On the other side, an opposite behaviour is observed when inspecting the spectral amplitude of man-made scenes, which reveals more their patterned texture as long as camera distance increases [27].

8 Experimental results

The extracted features for shot type classification feed two classifiers for further comparison: C4.5 decision trees [30] and Support Vector Machines (SVM) [12]. The adoption of these two learning algorithms is motivated by the fact that they constitute two representative samples among recent lines of research in machine learning techniques. SVM ensures high classification speed and accuracy, fair robustness to noisy data and irrelevant features; the downside is found in a slow learning process and in the difficulties of parameter handling and model comprehension by the user. Conversely, C4.5 builds a decision tree with a very intuitive procedure, allowing for a better understanding of the final model. Moreover it is generally fast in both learning and classification processes.

In particular, for both classifiers we provide confusion matrices and the following performance indices: *accuracy*, *specificity* and *F-measure* (with the details of *precision* and *recall*). Accuracy is the most common way of assessing classification results and it measures the proportion of true results (both true positives and true negatives). Specificity instead assesses how many negatives are correctly classified; such an indicator is important because in a real classification scenario a crucial objective is to avoid false positives. In addition we also report results in terms of precision and recall, and their aggregated form F-measure F_1, i.e., their weighted harmonic mean.

8.1 Data preparation

Our data corpus is composed of 3000 shots with starting resolution of $\mathcal{W} \times \mathcal{H}$, with $\mathcal{W} = 720$ and $\mathcal{H} = 480$, excerpted from 12 movies by different directors chosen from the Internet Movie Database (IMDb) [19], and filmed in a period which covers the last 30 years. Movie titles and the related genres can be found in Table 1.

Each movie is automatically divided into its shots, and for each shot the central frame is considered. Selected frames from all the movies constitute the starting data for our dataset. To build the actual dataset we use the following procedure: an algorithm randomly extracts frames from data which are manually annotated independently by authors[1] following the definitions given in Section 1: a shot is considered as a CU when it depicts human characters from the breast upwards, as a MS when it shows from the waist downward to include the whole body, while it is a LS when it privileges the background presence. The reason for choosing classes with a gap is that shots with close distance scores are not likely to have any distinguishing feature, and may merely be representing the noise in the whole peer-rating process. In case the extracted shot contains numerous actors with various postures at different distances from the camera, the closest actor facing the camera is considered as the reference for the labelling process.

In order to ensure balance among classes, the process ends when data corpus has gathered 1,000 Long shots, 1,000 Medium ones, and 1,000 Close-ups. Conversely the database is not balanced with respect to the presence of man-made and natural images. This reflects the intention of having a balance with respect to the main classification aim, which is the categorisation into shot types, while respecting in the data corpus the proportion between the presence of natural and man-made scenes in modern cinema.

For what concerns the prior probabilities of LS, MS, and CU in standard movies, these percentages vary significantly from one movie to another and from genre to genre. An automatic estimation on the complete movie database of Table 1 assesses the percentages of shot type presence as: 19% for Long shots, 35% for Medium, and 46% for Close-ups. Similar figures arise from the study carried out in [9] where we analyse 83 “great movie scenes” chosen to represent popular films from 1958 to 2009 (total duration of more than 3 h of video and 2311 shots). On the same databases, the proportion between natural and man-made scenes is estimated to be around 1/5.

Since the spectral feature vectors are differently computed depending on the nature of the image, we first need to pre-process shot images to distinguish between man-made and natural ones, thus obtaining two different datasets.

[1] The few labelling discrepancies are harmonised after discussion between the labellers.

Table 1 Film titles and their IMDb genre

No.	Movie title	Genre
1	Raiders of the lost ark	Action/adventure
2	War of the worlds	Action/adventure/drama
3	A beautiful mind	Biography/drama
4	All or nothing	Drama/comedy
5	Home	Documentary
6	Spring, summer, fall, winter... and spring	Drama
7	Eternal sunshine of the spotless mind	Drama/romance/sci-fi
8	Samaritan girl	Drama
9	Phone booth	Mystery/thriller
10	Seven swords	Action/fantasy
11	Once upon a time in the west	Western
12	All about my mother	Drama

The pre-classification step between natural and man-made shots can be implemented by using the method proposed by Torralba et al. in [36] which makes use of the spectral feature as a discriminant factor, as described in Section 7. With our database it allows for a correct categorisation of the 83% of images, in the specific 74% for LS, 85% MS, and 90% CU. Details about the training and classification procedures of this pre-processing step can then be found in the same work [36], since we have followed a similar training procedure. Although having remarkable performance, this algorithm is not error free. Therefore in order not to bias the two employed classification methods with possible errors in the training data, for the experimental phase we start with a perfect subdivision in the two datasets, man-made and natural.

For the whole annotated set of 3000 shots, the five features (histogram variance ratio A_σ, motion activity map ratio A_M, angular aperture for scene perspective α, detected face ratio A_F for absolute shot distance and global spectral amplitude Γ) are extracted to form the experimental data.

For both datasets, man-made and natural, half of the images is used for training the classifiers, while the classification task is performed on the second half of the dataset. Shots in the two halves are arranged following the *stratification* process [32], thus ensuring that in every fold each class comprises around half of the instances.

8.2 C4.5 decision trees: combined descriptors

Decision tree classifiers build "trees" by iteratively splitting the training set into subsets. At each node of the tree, the classifier chooses one of the data features that most effectively splits its set of samples, and the process is then iterated on the children nodes. This "divide and conquer" approach leads to a final tree-like structure in which each interior node corresponds to one of the input features, while each leaf represents a value of the target class given the values of the features represented by the path from the root to that leaf.

Different methods can be used for selecting the splitting descriptor, that is for deciding which of the features are the most relevant, so they can be tested near the root of the tree. In the C4.5 algorithm the default splitting criterion uses the concept of *information gain ratio*, based on the difference in entropy prior and subsequent to the splitting. Although this is usually a good measure for deciding the relevance of a feature, performance may be weak in domains with a preponderance of continuous features. The C4.5 algorithm handles this issue by creating at each step a threshold for the selected feature, splitting those samples whose values are above the threshold and those that are less than or equal to it. For insights on the algorithm, please refer to [31]. Our implementation uses the C4.5 algorithm working with a final pruning phase in an attempt to simplify the generated tree.

Confusion matrices for the three shot types are given in Table 2 for both man-made and natural images, respectively, while classification results obtained on the test datasets (man-made and natural) are shown in Table 3 in terms of accuracy, specificity and F-measure (with the details of precision and recall).

In both testing scenarios (natural and man-made) fair performance is achieved according to all the evaluation criteria. In addition to this, the fact that at each node the feature that best divides the training data is chosen points out the relevance of single features and their inter-relationships.

To understand the role of single features in the construction of the decision tree, that is their ability in dividing the training data, the interested reader can for example observe the first three levels of the decision trees depicted in Figs. 7 and 8 for man-made and natural images, respectively. From an inspection of both decision trees, it emerges the predominant role of two features: the one related to the presence of faces A_F, and the angular aperture of perspective lines α. While in the case of man-made (decision tree in Fig. 7) the presence of faces is the most significant descriptor, they switch their positions in natural images (Fig. 8) where the perspective aperture takes over. Other features intervene on deeper levels of the tree to support the decision process when a final categorisation is not reached on previous stages.

As a conclusive remark regarding C4.5 algorithm, even if other machine learning classifiers (such as SVM) might ensure higher classification accuracy, decision trees allow for deep understanding of the effectiveness of different feature descriptors as input for classifiers. This might be crucial for future improvements of the categorisation approach and its integration in a possible longer toolchain towards semantic analysis of fiction content.

8.3 Support Vector Machine: combined descriptors

Classification experiments involving SVM are here presented as complementary tests to those performed with C4.5, so that the combination of the two allows for a more complete view on the effectiveness of the proposed approach. SVM are supervised learning methods used for classification and regression, playing an increasing role in

Table 2 C4.5 classifier—confusion matrices (sum of each row is 1)

Image-type	Shot-type	LS	MS	CU
Man-made	LS	0.663	0.194	0.143
	MS	0.252	0.524	0.224
	CU	0.109	0.106	0.785
Natural	LS	0.818	0.106	0.076
	MS	0.342	0.316	0.342
	CU	0.365	0.095	0.540

Table 3 C4.5 classifier—shot type detection with (*up*) the combined set $\{A_\sigma, A_M, \alpha, A_F, \Gamma_m\}$ on man-made images, and (*down*) with the feature set $\{A_\sigma, A_M, \alpha, A_F, \Gamma_n\}$ on natural images

Image-type	Shot-type	Acc. (%)	Spec. (%)	F_1	Prec.	Rec.
Man-made	LS	**79.1**	83.5	0.619	0.579	0.664
	MS	**76.2**	86.2	0.567	0.616	0.524
	CU	**80.1**	81.4	0.780	0.774	0.785
Natural	LS	**74.4**	64.7	0.783	0.750	0.818
	MS	**80.3**	89.8	0.343	0.375	0.316
	CU	**77.8**	86.5	0.574	0.603	0.547

Accuracy, in bold, provides a summary of the performance

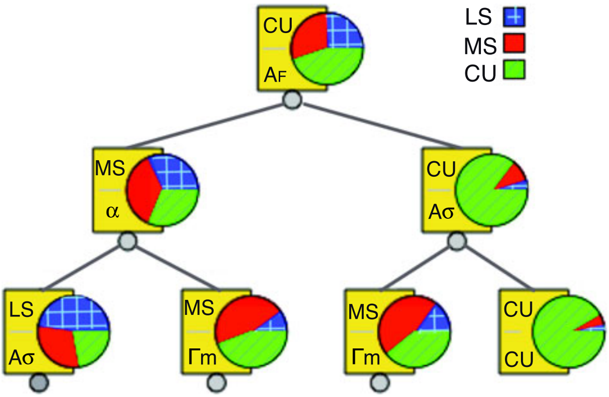
signal processing, pattern recognition and image analysis. The principle is that, given two classes of data which are not separable by a linear function, a SVM projects data into a higher dimensional space (via kernel representation), where the separation problem is solved by building an optimal separating hyperplane which maximises the functional margin.

For each dataset (man-made and natural) a multiclass SVM is trained on the combined feature set using the "one-against-one" approach [18], thus creating the models for the classification task. For each SVM, the penalty term C and parameter ξ of a standard RBF kernel $K(x, y) = \exp(-\xi \, \|x - y\|^2)$ are obtained performing cross validation just on the training set via a process of grid search to maximise cross validation accuracy. The best couples $(\hat{C}, \hat{\xi})$ are then used to train the two training sets and generate the final models.

Confusion matrices for the three shot types are given in Table 4 for both man-made and natural images, respectively; moreover, results obtained on the testing datasets in terms of accuracy, specificity, F-measure, precision and recall are shown in Table 5.

In both testing scenarios (natural and man-made) good performance is achieved according to all the evaluation criteria. It is also evident that scores using SVM with the combined descriptor sets are higher when compared to those obtained employing the C4.5 decision trees. This is coherent to what we expect from SVM, which in

Fig. 7 The first three levels of the decision tree for man-made images. In each node the splitting feature is shown, together with the majority class, and a pie diagram with the distribution of different shot types



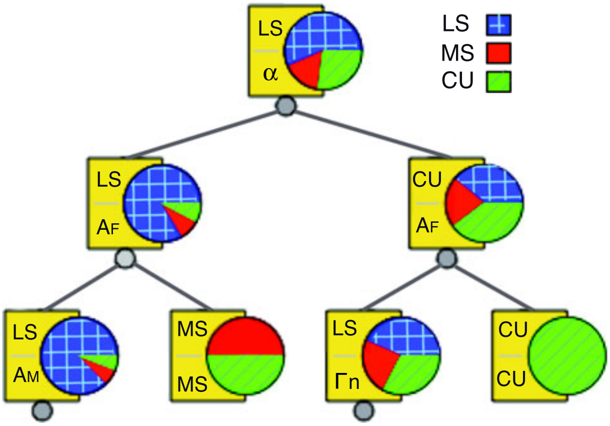


Fig. 8 The first three levels of the decision tree for natural images. In each node the splitting feature is shown, together with the majority class, and a pie diagram with the distribution of different shot types

general achieve a higher accuracy of classification than decision trees, especially when dealing with continuos or multi-dimensional features [22].

Despite the overall good performance with respect to accuracy, both classifiers show specificity values for natural LS which are significantly lower than for all the other classes. A tentative explanation could be that, in the case of shots depicting man-made scenes, apart from the foreground subject when present, the background often shows structured objects (such as buildings, etc.) which are helpful for a correct classification of the camera distance category. Conversely, shots of natural scenes provide generally fewer background or structural information on whose basis to drive a decision on the shot type, a situation which is indeed very typical in Long shots, where the distance of the camera from the scene may "hide" important revealing details. For this reason, a remarkable number of natural MS and CU are misclassified as LS. This interpretation, enforced by the confusion matrices on natural shots for both classifiers (Tables 2 and 4), might explain the over-classification and the consequent low specificity of Long shots in natural scenes (Tables 3 and 5), as well as the very low specificity of the spectral feature on Long shots of natural images for SVM which will be given in Table 6.

8.4 Support Vector Machine: single descriptors

It is evident that scores which are obtained using all combined features cannot be outperformed using only individual features. In fact, although certain individual

Table 4 SVM classifier—confusion matrices (sum of each row is 1)

Image-type	Shot-type	LS	MS	CU
Man-made	LS	0.692	0.198	0.110
	MS	0.192	0.628	0.180
	CU	0.081	0.070	0.849
Natural	LS	0.951	0.022	0.027
	MS	0.447	0.461	0.092
	CU	0.384	0.040	0.576

Table 5 SVM classifier—shot type detection with (*up*) the combined set $\{A_\sigma, A_M, \alpha, A_F, \Gamma_m\}$ on man-made images, and (*down*) with the feature set $\{A_\sigma, A_M, \alpha, A_F, \Gamma_n\}$ on natural images

Image-type	Shot-type	Acc. (%)	Spec. (%)	F_1	Prec.	Rec.
Man-made	LS	**82.8**	87.5	0.672	0.653	0.692
	MS	**80.8**	88.4	0.660	0.695	0.628
	CU	**85.1**	85.2	0.836	0.824	0.849
Natural	LS	**79.6**	59.2	0.840	0.754	0.951
	MS	**88.8**	97.2	0.574	0.761	0.461
	CU	**85.6**	95.9	0.682	0.837	0.576

Accuracy, in bold, provides a summary of the performance

features might be effective even if taken alone, their inter-combination in a collaborative fashion almost certainly improves the classification performance. However it is still interesting to understand the ability of each single feature in distinguishing the shot type, beyond the analysis already presented in Section 8.2 on the decision trees.

To this aim, in this second part of the experiment the features are tested individually using SVM classifiers, so that to assess their utility in shot type classification. Results for single features are reported in Table 6.

Specifically, results show that the classifier related to the presence of faces (A_F) achieves good performance, according to all of the considered evaluation criteria. This is evident, since when a face is correctly detected it provides a clue of absolute distance, being an element with a well defined dimensional scale. When no human beings are depicted, the shot is classified as Long, while theoretically it could be also a CU of a generic object.

Even if the probability of having a face in a scene is high due to the human centrality to the narrative perspective, it is yet difficult producing an accurate (i.e., automatic) estimation of the face presence in movie shots. As an example, Long

Table 6 Shot type classification results obtained with SVM using single features $\{A_\sigma\}$, $\{A_M\}$, $\{\alpha\}$, $\{A_F\}$, $\{\Gamma_m\}$ and $\{\Gamma_n\}$

Shot-type	**Acc.(%)**	Sp.(%)	F_1	**Acc.(%)**	Sp.(%)	F_1
	A_σ (colour)			A_M (motion)		
LS	**67.8**	97.9	0.143	**48.7**	36.6	0.487
MS	**59.5**	56.2	0.518	**62.6**	80.3	0.303
CU	**68.8**	68.2	0.602	**63.1**	88.3	0.203
	α (geometry)			A_F (face)		
LS	**66.6**	60.6	0.611	**66.5**	60.0	0.613
MS	**66.8**	85.2	0.362	**72.6**	82.6	0.553
CU	**64.7**	77.6	0.431	**76.3**	93.9	0.545
	Γ_m (spectral—man-made)			Γ_n (spectral—natural)		
LS	**75.9**	96.9	0.237	**57.8**	2.0	0.730
MS	**68.5**	79.0	0.453	**83.8**	100.0	0.026
CU	**67.0**	50.7	0.703	**73.9**	100.0	0.047

Accuracy, in bold, provides a summary of the performance shots sometimes show human figures from the distance. In this case, human faces are present, but due to their reduced dimensions, they are hardly detectable. On the other hand, shots without people are often establishing shots (i.e., again LS). Therefore we would tend to state that no human faces are actually present (or "detectable") in our database of Long shots. Conversely, a large majority (>85%) of Close-ups actually contains human faces, since they are mostly used to focus on human reactions. Eventually, for Medium shots we estimate the presence of human faces to be in the interval between 50% and 55%.

The classifier trained with the angular aperture of perspective lines (α), as well as the one obtained by the analysis of colour intensity distribution over local regions (A_σ), have good overall performance, even if they suffer from unbalance between precision and recall, highlighted by low values of F_1 for some classes. From this perspective, the classifier trained with the motion activity maps (A_M) is the less performing one among those in the pixel domain, even though it has good accuracy for two classes of data (MS and CU).

Regarding the spectral feature, two different runs, first on man-made and then on natural images are carried out. In the last row of Table 6, classification results on man-made images are on average higher than those obtained in the pixel domain, and comparable to the highest ones achived by the face descriptor. For the natural set instead, as commented above, a closer look at the $F1$ indicator and at the specificity reveals that the spectral descriptor, when considered alone, is unable to properly capture the characteristics of the natural dataset.

In ultimate analysis, despite the different nature of the two classifiers, the dominant role of single descriptors A_F and α here described for SVM, is also consistent with the analysis previously illustrated on decision trees.

9 Conclusions

In this work, we propose a method for estimating the distance between camera and the filmed subject without recovering the 3D structure of the scene. By investigating five features which provide clues about the shot type, we classify movie shots into Long, Medium, and Close-ups. The first feature accounts for colour intensity distribution on local regions in frames; the second employs Motion activity maps to estimate the occupancy of the frame space by moving foreground objects. The third method relies on the 2D geometry of the scene, by measuring the angular aperture of perspective lines. Fourth, when faces are present, their dimensions provide an indirect measure of the absolute distance between the camera and the filmed subject. Finally, the decay of the spectral amplitude of the image transform provides information about the dominant structure and scale of the scene, for both natural and man-made images. In the experimental phase, using C4.5 decision trees and Support Vector Machines, we combine all extracted features to achieve high classification performance according to all considered evaluation criteria.

9.1 Future applications

Since camera distance deeply affects the emotional involvement of the audience and the process of identification of viewers with the movie characters [1], extending the idea further, the study of inter-shot relationships can pave the way for investigating the affective reactions of users to different patterns of shot types. On the basis of the proposed shot type classifier, the work in [9] already investigates the use of camera distance in famous movie scenes, highlighting the relations between the employed shot types and the affective responses by a large audience. Obtained results suggest that patterns of shot types constitute a key element in inducing affective reactions in the audience, with strong evidences especially on the arousal dimension. These findings are therefore applicable to support systems for media affective analysis, and to better define emotional models for video content understanding. Moreover, when shooting dialogues, directors often follow film grammar rules suggesting the usage of specific patterns of shot types [1], which could be easily detected thanks to the proposed techniques. The long-term aim is to integrate the shot type classifier in a longer toolchain towards semantic analysis of fiction content, with a particular attention to the emotional reactions of the audience.

Another envisaged study based on this work aims at the automatic characterisation of the psychological role of characters in movies. For example the massive use of close-ups focusing on characters' emotional feelings, beyond boosting the process of identification of viewers with the film characters, is useful to sketch psychological relationships between characters. In addition to this, the use of certain shot types such as the "over-the-shoulder" shot when two characters are having a discussion, is often employed when the director wants to stress a situation of psychological dominance of one characters over the other. With these premises, shot type classification might be exploited in the context of video story-telling [29] for the automatic composition or recombination of video shots.

Eventually, repositories of shot annotated with their related shot type could be useful for new forms of emerging creativity such as the practice of combining multiple audiovisual sources into a derivative work (known as video mashup) whose semantics could be very different compared to the one of the original videos. Automatic or semi-automatic tools (such as that described in [8]) able to combine shots according to filmic grammar rules could undoubtedly benefit of such annotated shot content.

Acknowledgement Figures 7 and 8 are modified versions of decision tree graphs obtained using the Orange software, available at http://orange.biolab.si.

References

1. Arijon D (1991) Grammar of the film language. Silman-James Press
2. Barrow H, Tenenbaum J (1981) Interpreting line drawings as three-dimensional surfaces. Artif Intell 17(1–3):75–116
3. Benini S, Canini L, Leonardi R (2010) Estimating cinematographic scene depth in movie shots. In: Proceedings of the IEEE International Conference on Multimedia & Expo (ICME). Singapore
4. Benini S, Xu LQ, Leonardi R (2005) Using lateral ranking for motion-based video shot retrieval and dynamic content characterization. In: Proc. of CBMI. Riga, Latvia
5. Bordwell D, Thompson K (1997) Film art: an introduction. McGraw-Hill
6. Bradski G (2000) The OpenCV library. Dr. Dobb's Journal of Software Tools
7. Brooks MJ (1989) Shape from shading. MIT Press, Cambridge, MA, USA
8. Canini L, Benini S, Leonardi R (2010) Interactive video mashup based on emotional identity. In: Proceedings of the 2010 European Signal Processing Conference (EUSIPCO). Aalborg, Denmark

9. Canini L, Benini S, Leonardi R (2011) Affective analysis on patterns of shot types in movies. In: Proceedings of the 7th international symposium on Image and Signal Processing and Analysis (ISPA). Dubrovnik, Croatia
10. Cantoni V, Lombardi L, Porta M, Vallone U (2001) Qualitative estimation of depth in monocular vision. In: Proc. of IWVF. Springer, London, UK, pp 135–144
11. Cherif I, Solachidis V, Pitas I (2007) Shot type identification of movie content. In: Proceedings of international symposium on signal processing and its applications. Sharjah, United Arab Emirates
12. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
13. Duan LY, Xu M, Yu XD, Tian Q (2002) A unified framework for semantic shot classification in sport videos. In: Proc. of ACM MM. ACM, New York, NY, USA, pp 419–420
14. Duda RO, Hart PE (1972) Use of the hough transformation to detect lines and curves in pictures. Commun ACM 15(1):11–15
15. Ekin A, Tekalp AM (2003) Robust dominant color region detection and color-based applications for sports videos. In: Proc. of ICIP'03. Barcelona, Spain, pp 1025–1028
16. Hanjalic A (2006) Extracting moods from pictures and sounds. IEEE Signal Process Mag 23(2):90–100
17. Hoiem D (2007) Seeing the world behind the image: spatial layout for 3d scene understanding. Ph.D. thesis, Robotics Institute, Carnegie Mellon Univ., Pittsburgh, PA
18. Hsu CW, Lin CJ (2002) A comparison of methods for multi-class support vector machines. IEEE Trans Neural Netw 13(2):415–425
19. Internet Movie Database (IMDb) http://www.imdb.com/. Accessed 2 May 2011
20. Jeannin S, Divakaran A (2001) Mpeg-7 visual motion descriptors. IEEE Trans Circuits Syst Video Technol 11(6):720–724
21. Keller JM, Crownover RM, Chen RY (1987) Characteristics of natural scenes related to the fractal dimension. IEEE Trans Pattern Anal Mach Intell 9(5):621–627
22. Kotsiantis SB (2007) Supervised machine learning: a review of classification techniques. Informatica 31:149–268
23. Kurita T, Otsu N, Abdelmalek N (1992) Maximum likelihood thresholding based on population mixture models. Pattern Recogn 25(10):1231–1240
24. Matessi A, Lombardi L (1999) Vanishing point detection in the hough transform space. In: Proc. of Euro-PAR '99. Springer, London, UK, pp 987–994
25. Monaco J (1981) How to read a film. Oxford University Press, New York
26. Nagai T, Naruse T, Ikehara M, Kurematsu A (2002) Hmm-based surface reconstruction from single images. In: Proc. of ICIP'02. Rochester, NY, USA, pp. 561–564
27. Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. Int J Comput Vis 42(3):145–175
28. Palmer SE (1999) Vision science-photons to phenomenology. MIT Press, Cambridge, MA
29. Porteous J, Benini S, Canini L, Charles F, Cavazza M, Leonardi R (2010) Interactive storytelling via video content recombination. In: Proceedings of ACM conference on multimedia (ACM MM). Florence, Italy
30. Quinlan JR (1993) C4.5: programs for machine learning (Morgan Kaufmann Series in Machine Learning), 1 edn. Morgan Kaufmann
31. Quinlan JR (1996) Improved use of continuous attributes in C4.5. J Artif Intell Res 4:77–90
32. Refaeilzadeh P, Tang L, Liu H (2009) Cross validation. In: In encyclopedia of database systems
33. Salt B (2006) Moving into pictures. More on film history, style, and analysis. Starword, London
34. Shimshoni I, Moses Y, Lindenbaum M (2000) Shape reconstruction of 3d bilaterally symmetric surfaces. Int J Comput Vision 39(2):97–110. doi:10.1023/A:1008118909580
35. Super BJ, Bovik AC (1995) Shape from texture using local spectral moments. IEEE Trans Pattern Anal Mach Intell 17(4):333–343
36. Torralba A, Oliva A (2002) Depth estimation from image structure. IEEE Trans Pattern Anal Mach Intell 24(9):1226–38
37. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Proc. of CVPR
38. Wang HL, Cheong LF (2009) Taxonomy of directing semantics for film shot classification. IEEE Trans Circuits Syst Video Technol 19:1529–1542
39. Xie L, Chang SF, Divakaran A, Sun H (2002) Structure analysis of soccer video with hidden markov model. In: Proceedings of ICASSP'02. Orlando, Florida, USA
40. Zeng W, Gao W, Zhao D (2002) Video indexing by motion activity maps. In: Proc. of ICIP. Rochester, USA

Luca Canini recieved his MSc in Telecommunications Engineering (cum laude) at the University of Brescia with a thesis which won a prize granted by the Italian Marconi Foundation. He is currently a PhD candidate in the same university. During his PhD studies he has been a visiting student at the IVE Lab, University of Teesside (UK) and at the DVMM Lab, Columbia University (USA).

Sergio Benini received his MSc degree in Electronic Engineering (cum laude) at the University of Brescia in 2000 with a thesis which won a prize granted by Italian Academy of Science. Between May 2001 and May 2003 he has been working in Siemens Mobile Communication R&D, on mobile network management projects. He received his Ph.D. degree in Information Engineering from the University of Brescia in 2006, working on video content analysis. During his Ph.D. studies, between September 2003 and September 2004 he has conducted a placement in British Telecom Research, Ipswich, U.K. working in the “Content & Coding Lab”. He is currently an Assistant Professor in the Telecommunications group of DII at the University of Brescia, Italy.

Riccardo Leonardi has obtained his Diploma (1984) and Ph.D. (1987) degrees in Electrical Engineering from the Swiss Federal Institute of Technology in Lausanne. He spent one year (1987–88) as a post-doctoral fellow with the Information Research Laboratory at the University of California, Santa Barbara (USA). From 1988 to 1991, he was a Member of Technical Staff at AT&T Bell Laboratories, performing research activities on image communication systems. In 1991, he returned briefly to the Swiss Federal Institute of Technology in Lausanne to coordinate the research activities of the Signal Processing Laboratory. Since February 1992, he has been appointed at the University of Brescia to lead research and teaching in the field of Telecommunications. His main research interests cover the field of Digital Signal Processing applications, with a specific expertise on visual communications, and content-based analysis of audio-visual information. He has published more than 100 papers on these topics. Since 1997, he acts also as an evaluator and auditor for the European Union IST and COST programmes.