Over-the-Shoulder Shot Detection in Art Films

M. Svanera, S. Benini, N. Adami, and R. Leonardi
DII, University of Brescia, Brescia, Italy
m.svanera005@unibs.it

A. B. Kovács
Film Department ELTE University, Budapest, Hungary
kovacs.andras.balint@btk.elte.hu

Abstract—The ability to characterize a film, in terms of its narrative and style, is becoming a necessity especially for developing personal video recommendation systems to better deliver on-demand Internet streaming media. Among the set of identifiable stylistic features which play an important role in the film’s emotional effects, the use of Over-the-shoulder (OtS) shots in movies is able to convey a big dramatic tension on the viewers. In this work we propose a methodology able to automatically detect this kind of shots by combining in a SVM learning scheme some state-of-the-art human presence detectors, with a set of saliency features based on colour and motion. In the experimental investigation, the comparison of obtained results with manual annotations made by cinema experts proves the validity of the framework. Experiments are conducted on two art films directed by Michelangelo Antonioni belonging to his famous “tetralogy on modernity and its discontent”, one in shades of gray (L’avventura, 1960), and the other in colour motion (Il deserto rosso, 1964).

I. INTRODUCTION

In film-making, the particular usage of the camera and its spatial relation to the subject greatly affects the narrative power of a shot. In this way directors manage to provide emphasis on key passages of the filmed scene and convey emotions to the viewer. For this reason an automatic detection of Over-the-shoulder (OtS) shots might constitute an important marker of a film’s stylistic and emotional character, often typical of an individual author or a period [1].

Among the standard film techniques in composing shots and choosing camera angles, Over-the-shoulder shot contains at least two persons, A and B, facing each other, with the camera shooting over the shoulder of one of them (A). Some examples of OtS as conceived by famous director Michelangelo Antonioni are given in Figure 1. Most often this type of composition is used when two characters have a conversation and OtS is part of the shot-reverse shot alternation where the camera angle is very close to character A’s point of view. According to cinematography three features characterise OtS: firstly, character A is always seen from the back. Secondly, this character can only be seen from the shoulder up occupying not more than a limited portion of the frame. Thirdly, character A is always closer to the camera than character B.

The effect of OtS is always a bigger dramatic tension than a simple Point-of-View (PoW) shot where the subject of the PoW cannot be seen. This tension is due to several factors: first, the difference between the dimensions of the two characters (character A is much bigger than character B, and sometimes in an apparently higher position, such as in Figure 1-d) which creates visual tension. Second, character A is shown as actively present in the space, since the viewer cannot see his/her face, he/she becomes somewhat unpredictable, and represents a potential point of attack (e.g., see Figure 1-a). Third, the relationship between the two characters is more tangible: a simple Close-up of the face of character A provokes identification with him/her, while OtS divides this effect between character A and B, and makes it ambiguous (as in Figure 1-b). All of these factors enhance the dramatic tension of the narrative.

A. Related work

Some works in literature have dealt with shot type classification in movies. The two approaches presented in [2] and [3] define human body-based rules to extract the shot scale, i.e. the apparent distance of the camera from the main subject of the scene. The work in [4] uses camera distance from the focus of attention as a feature to distinguish contextual-tracking and focus-tracking shots on a significant data corpus. A taxonomy based on camera motion is used in [5] for classifying cinematographic shots into eight categories: Aerial, Bird eye, Crane, Dolly, Establishing, Pan, Tilt, and Zoom.

Not many works, however, have dealt with the specific detection of OtS in video material: apparently this identification is still a difficult problem when working only with colour/gray-scale image features. Although many researches have made attempts to detect human characters from behind, like in [6], [7] or [8], only few works demonstrated to be able to detect the shot type in analysis, and to a limited extent.

In [9] Hoai et al., using an upper-body classifier, aimed at recognizing different configurations of one or more people in edited TV material, and to discriminate between different human interaction classes, with a good average precision. By

Fig. 1. Examples of Over-the-Shoulders in art films. Colour images (a-b) are taken from Il deserto rosso (1964). Gray-scale images (c-d) are taken from L’avventura (1960). Both movies are directed by Michelangelo Antonioni.
imposing some constraints on locations, scales, and configurations of people which are highly specific to TV videos, they limit the application of this methodology to TV material only, preventing its application for example to the higher variety of situations which can be encountered in movies and especially in art movies.

Xu et al. in [10] and [11] provide a system for classifying shot scales, including less traditional types such as Over-the-shoulder and Foreground shots (here called Cut-away). The method is based on a context saliency map generated by removing redundancy from contrast saliency and incorporating geometry constrains. Saliency is further combined with color and texture to generate feature vectors to feed an SVM for training and classification, with fair performance on OtS class (F-measure: 0.51%).

In [12] authors proposed an OtS detector applied on 3D (stereo) video. To solve the problem they exploited the information extracted from a disparity map estimated by the method in [13]. The classification accuracy of each class accounts for 93.62% for non-OtS, and 86.54% for OtS, for a mean accuracy of 91.65%. However currently only few movies are available as 3D data. Since most of films and TV series are still produced in 2D, this requires a new approach to the problem.

B. Paper aim and organization

In the work described in [14] we proposed a framework to estimate the shot scale distribution of art movies, limiting the analysis to three shot scales: Long-shots, Medium-shots and Close-ups, respectively. Since Kovács in [1] observed a systematic variation of shot scale distribution patterns in films by Michelangelo Antonioni (including OtS), being able to characterise a movie on all available shot scales could be an important identifier not only of the film, but also of the style of a particular director. For this reason, in this paper we extend the previous analysis on shot scales and capture the presence of Over-the-Shoulder shots by proposing a methodology and a processing pipeline for their automatic detection.

As shown in Figure 2, this objective is pursued by combining state-of-the-art human presence detectors with a set of saliency features based on colour and motion. Despite in OtS shots there are usually at least two persons facing each other, with the camera shooting over the shoulder of one of them, even most advanced human presence detectors often fail in revealing both actors. As a matter of fact, the person filmed from behind is undoubtedly the most characterising OtS feature, and at the same time, the most difficult to automatically capture. This work demonstrates the determinant role played by the set of saliency features in improving detection performance of such difficult shot material. The test phase involves two entire art movies, one in colour and one in grey-scale, belonging to the Antonioni’s “tetralogy on modernity and its discontents”.

The paper is organized as follows. Section II describes the set of adopted human presence detectors (i.e., pedestrian, upper-body, and face detection). Section III provides an overview of employed features based on saliency. Section IV presents our experimental results, while in Section V conclusions are finally drawned.

Fig. 2. System workflow for OtS shots detection, where human presence detectors are combined with saliency based features to feed a SVM learner.

II. DETECTION OF HUMAN PRESENCE

Apart from few people-less movies, the presence of human figure is central to modern cinematography. In the following we present few advanced techniques to detect the presence of characters in various poses, trying also to consider the frequent case when actors are partially (or almost totally) occluded. The use of multiple human presence detectors is justified by the fact that in OtS actors may be shot at different distance.

A. Pedestrian detection

When actors are filmed in long scale, the detection of the full body, beyond providing a direct measure of the scale itself, carries an important clue for the classification of OtS shots. The employed detector is based on mixtures of multi-scale deformable part models by Felzenszwalb et al. [15] trained for the specific task of pedestrian detection. Once detected all bodies in the image, a pedestrian map is obtained by labelling in the image the correspondent bounding boxes. An example of the detection mask is shown in Figure 3-b in transparency.

B. Upper-body detection

In [9] Hoai and Zisserman propose an upper-body detector for spotting particular people configurations in TV material. Modifying a version of the Deformable Part Model (DPM) proposed by Felzenszwalb et al. [15], they exploit the fact that the locations, scales, and configurations of people in TV video are constrained to few configurations. Although art movies are less adherent to standard techniques of composing shots and choosing camera angles than TV material, we adopt this detection algorithm in the constrained configuration of having two main characters in the image frame, which is the most frequent case in OtS shots. Once located the presence of upper bodies, a binary map is produced in correspondence of the related bounding boxes, as shown in Figure 3-c in transparency.

C. Face detection

According to the provided definition, an OtS should contain at least one face belonging to the frontal character. To ensure
higher recall rates even on non-perfect frontal views, the method proposed by Zhu and Ramanan in [16] is very effective at capturing global elastic deformation of faces. This algorithm employs a tree structure composed by a set of parts (e.g., HoG descriptor) modelling every facial landmark as a part, and uses global mixtures to capture topological changes. Although this method provides also with the pose estimation and landmark positions, in order to obtain a face location map we only label the area covered by detected faces in the image. An example of the obtained face map is shown in transparency in Figure 3-d.

![Image](image_url)

Fig. 3. Features extracted from a) an original image from *Il deserto rosso* (1964) di Michelangelo Antonioni. Human presence detection (green): b) pedestrian detection, c) upper-body detection, d) face detection. Saliency features (yellow): e) multi-scale contrast, f) colour-intensity histogram, g) center-surround histogram, h) colour-spatial distribution, and i) motion map (normalized for visualization purposes).

### III. SALIENCY FEATURES

The following five saliency characteristics based on colour distribution and amount of motion in the image hint at the presence of persons in the scene, and want to be an effective and complementary support to the detection techniques described in Section II. In fact approaches relying only on the presence of human faces, or the detection of other parts of the human body (as those described in Section II) might fail in the task of OtS detection. For example, it is often the case that the person facing away from the camera is cut off just behind the ear. Even in the case when a bigger portion of the shoulder, or of the back, of the person facing the frontal subject is included, no effective solution currently exists for detection, so that complementary features based on saliency are needed for improving detection performance.

#### A. Multi-scale contrast

Since characters pictured in a movie frame are to be considered the salient objects of the scene, it is beneficial to describe the presence of actors in terms of visual saliency. Among commonly used features, contrast is one of the most adopted for attention detection since it simulates the human visual receptive fields [17]. An effective way to analyse saliency is to make use of multi-scale contrast as proposed in [18]. In particular multi-scale contrast feature $f_{\text{msc}}(x, y)$ is defined as a linear combination of contrasts in the Gaussian image pyramid:

$$f_{\text{msc}}(x, y) \propto \sum_{l=1}^{6} \sum_{(x', y') \in N(x, y)} \| I_l(x, y) - I_l(x', y') \|^2$$

where $I_l$ is the $l$th-level image in the pyramid and $N(x, y)$ is a 9×9 window, adding *de facto* the contrast found at six different scales and normalizing the results to the range $[0, 255]$.

Analysing the scene by a multi-scale approach based on visual saliency we are able to well capture the presence of salient characters in the OtS scene, in particular highlighting high contrast boundaries, while giving lower scores to the homogenous regions inside the salient characters, as shown in the obtained multi-scale contrast image shown in Figure 3-e. This feature works particularly well in OtS shots when the actors cover a wide area of the frame, as in the example of Figure 4, thus becoming the prevalent objects of the scene.

![Image](image_url)

Fig. 4. a) A frame from *L’avventura* (1960) with b) the corresponding multi-scale contrast image.

#### B. Colour-intensity histogram

The previous approach to the problem of saliency works well when both subjects are shot in a close to medium scale. When the frontal character is further away from the camera, it is more advantageous to consider a local approach to the problem of saliency, able to distinguish, for example, the situation in Figure 1-a from the one Figure 1-b. If fact, when looking to a picture, distant objects are not as sharp as those of foreground, due to the diffusion and diffraction of light in an opaque medium [19]. Intensity images become then more blurred and uniform as the distance from the camera increases; on the contrary, in areas with edges are sharper (i.e., nearer to the camera) intensity levels are more scattered. For this reason, a local approach to saliency better captures the subject details when he/she is pictured far away from the camera. As we proposed in [20], the *colour-intensity histogram* is obtained measuring the second order statistics of local image histograms on the corresponding intensity image. A local histogram is computed over $N(x, y)$, a squared sliding window centered on pixel $(x, y)$ and scanning the intensity image $I$. Indicating with $f(g, N)$ the number of pixels in the window $N(x, y)$ whose gray level is equal to $g$, the average gray level of the
To handle varying aspect ratios of salient characters, by the expression of a distance with larger dimensions (width and height), is easily identifiable (defined as the point of junction between the diagonals) but colour distribution) and a rectangle [18]. Supposing that the relevant character is confinable within regional salient feature, such as the be distinguished from its surrounding context by employing a C. Center-surround histogram colour-intensity histogram image.

Fig. 5. a) A frame from Il deserto rosso (1964) with b) the corresponding colour-intensity histogram image.

C. Center-surround histogram

When actors gain a large extent in the frame, they can be distinguished from its surrounding context by employing a regional salient feature, such as the center-surround histogram [18]. Supposing that the relevant character is confinable within a rectangle $R$, the visual differences between $R$ (in terms of colour distribution) and a rectangle $R_s$ with the same center (defined as the point of junction between the diagonals) but with larger dimensions (width and height), is easily identifiable by the expression of a distance $\chi^2$ between the corresponding histograms of RGB colour (or gray-scale), i.e.:

$$\chi^2(R, R_s) = \frac{1}{2} \sum_{i \in \{R,G,B\}} \frac{(R_i - R_s)^2}{(R_i + R_s)}$$

To handle varying aspect ratios of salient characters, five templates with different aspect ratios are used {0.5, 0.75, 1.0, 1.5, 2.0}. The most distinct rectangle $R^*(x, y)$ centered at each pixel $(x, y)$ is found by varying the size and aspect ratio, so that:

$$R^*(x, y) = \arg \max_{R(x,y)} \chi^2(R(x,y), R_s(x, y))$$

The center-surround histogram feature $f_{csd}(x, y)$ is defined as a sum of spatially weighted distances

$$f_{csd}(x, y) \propto \sum_{(x', y')} \omega(x,y)(x', y') \chi^2(R^*(x', y'), R_S(x', y'))$$

with $(x', y')(x, y) \in R^*(x', y')$, where $R^*(x', y')$ is the rectangle centered at $(x', y')$ and containing the pixel $(x, y)$. The weight $\omega(x,y)(x', y')$ is a Gaussian falloff weight with variance $\sigma^2(x', y')$, which is set to one third of the size of $R^*(x', y')$. The result is a center-surround histogram image as shown in Figure 3-g. In Figure 6 another example of salient characters well located by the center-surround histogram feature is shown.

Fig. 6. a) A frame from Il deserto rosso (1964) and b) the corresponding Center-surround histogram image.

D. Colour-spatial distribution

Since its introduction, the use of colour has been a powerful storytelling device in films. Colour usage can be intrinsic to the story narrative, for example differentiating story strands by using different colour palettes, or more symbolic, for example representing different ideas or characters’ qualities with respect to the surrounding scene (see for example Monica Vitti’s green coat in Figure 3-a). Hence the global spatial distribution of a specific colour, as proposed in [18], is a suitable descriptor for capturing the global relevance of a scene object or character. The idea is that the smaller a colour variance is, the higher probability the colour belongs to a salient object (or equivalently, the wider presence of a colour in the image, the less probably a salient object contains this colour). In the specific the spatial distribution of a specific colour (or of a specific gray level in single intensity images) is computed as the spatial variance of the colour itself. All colours in the image $I$ are first represented by Gaussian Mixture Models (GMMs) and each pixel $(x, y)$ is assigned to a colour component $c$ with a probability:

$$p(c|I(x,y)) = \frac{\omega_c N(I(x, y)|\mu_c, \Sigma_c)}{\sum_c \omega_c N(I(x, y)|\mu_c, \Sigma_c)}$$

Then, the horizontal $V_h(c)$ and vertical $V_v(c)$ variances of the spatial position for each colour component are computed for each colour component $c$, and summed as $V(c) = V_h(c) + V_v(c)$. The colour-spatial distribution feature $f_{csd}(x, y)$ is finally defined as the weighted sum:

$$f_{csd}(x, y) \propto \sum_c p(c|I(x,y)) \cdot (1 - V(c))$$

which produces a corresponding colour-spatial distribution image to the range [0, 255] as the one shown in the example of Figure 3-h, where the green coat of Monica Vitti stands out with respect to the other objects in the frame. Another example where the colour-spatial distribution clearly highlights the presence of salient characters is presented in Figure 7.

Fig. 7. a) A frame from Il deserto rosso (1964) with b) the corresponding colour-spatial distribution image.
E. Motion map

While previous techniques for detecting salient objects (i.e., actors) are based on colour, another criterion for saliency estimation is derived from a motion descriptor able to characterise the perceived activity of motion in a video segment, as well as its unique spatial distribution. In particular we extract a Motion map as in [20] analyzing motion of a video segment from the image plane along its temporal axis. Motion maps are extracted from predicted frames of the MPEG-4 compressed stream and are gray-scale images showing the amount of motion which has been accumulated along one second of analysis. In particular, from each video segment $S$ of one second with frame dimension $W \times H$ a corresponding motion map $f_{mm}$ with same dimensions is extracted. Each motion map is made up of $\frac{W}{Q} \times \frac{H}{Q}$ macroblocks of $Q \times Q$ identical pixels, where the value of $Q$ depends on the adopted codec - typical values are $Q = 4, 8, 16$. The value of each pixel $(x, y)$ of $f_{mm}$ is the normalised numeric integral, computed over all predicted frames $f_p$ of the shot $S$, of the magnitudes of motion vectors $m_v(B_{i,j})$ associated to the macroblock $B_{i,j}$ containing pixel $(x, y)$, that is:

$$f_{mm}(x, y) \propto \frac{1}{\# f_p} \sum_{f_p \in S} |m_v(B_{i,j})|_{f_p} \text{ s.t. } (x, y) \in B_{i,j}$$

Therefore in a motion map, single pixel intensities measure the amount of motion undergone by the corresponding $Q \times Q$ macroblock $B_{i,j}$ averaged over the video segment duration, and normalised to a 8-bit over the entire movie. The utility of a motion map is twofold: as shown in Figure 3-i, on the one hand it provides a clue about presence and rough dimensions of salient moving objects, while, on the other hands, it indicates if motion saliency is spread across many regions or restricted to a large one, providing a view of its spatial distribution. Another example where the motion map clearly highlights the presence of salient characters is presented in Figure 8.

![Fig. 8. A frame from L'avventura (1960) and b) the related motion map.](image)

IV. EXPERIMENTAL RESULTS

To evaluate the proposed method in terms of its classification ability, two entire movies directed by Michelangelo Antonioni, namely L'avventura (1960) and Il deserto rosso (1964), have been considered. Art films, due to the variety of experimental aesthetic situations, the richness of the scene composition, and the presence of unconventional or highly symbolic content [21], may be considered among the most challenging material for automatic movie analysis. Conversely common TV material and mass commercial movies usually adopts more standard techniques of shot composition and choice of camera angles.

In order to cover different types of colour material, L'avventura (1960) is a “black-and-white” movie, meaning that pictured frames are in shades of gray, while Il deserto rosso (1964) is a colour motion film. All video material has been manually annotated by a cinema expert, who classified each frame - extracted at 1 fps - either with “Over-the-Shoulder” (OoS) label or “non-OoS” one. The cardinalities of OoS and non-OoS classes are reported in Table I for both movies.

<table>
<thead>
<tr>
<th>Movie title</th>
<th>Year</th>
<th># frames</th>
<th># OoS</th>
<th># non-OoS</th>
</tr>
</thead>
<tbody>
<tr>
<td>L'avventura</td>
<td>1960</td>
<td>8435</td>
<td>223</td>
<td>8212</td>
</tr>
<tr>
<td>Il deserto rosso</td>
<td>1964</td>
<td>6537</td>
<td>193</td>
<td>6344</td>
</tr>
</tbody>
</table>

Feature vectors are extracted at the time resolution of 1 second, according to the workflow described in Figure 2, for both whole movies, meaning that no frame nor second has been discarded by the analysis. All the features, but the motion map, are computed on single image frames extracted at the time resolution of 1 fps. The motion map is again produced each second, but uses all movie frames (25 fps) as inputs to compute the related feature vector. Once extracted, all features are resized to smaller dimensions ($100 \times 84$ pixels) to reduce the dimensionality of the combined feature vector. The final feature vector, which feeds a SVM classifier, has total dimensions of $107520 (160 \times 84 \times 8)$, where non-zero values are scaled to the range $[0, 1]$.

With respect to the SVM classification stage, the two classical phases - training and testing - have been first performed on single movies. For each movie, a balanced training has been carried out by randomly extracting, for both classes, a number of frames corresponding to the 75% of the frames of the least populous class (OoS), for a total of 334 ($167 + 167$) training samples for L'avventura and 288 ($144 + 144$) samples for Il deserto rosso. The remaining frames, 8101 for L'avventura, and 6249 for Il deserto rosso, have been used for the testing phase, for a total number of almost 15000 testing images. The process of extracting training samples, the training phase and the testing phase, have been then repeated 20 times in order to guarantee the invariance and stability of the analysis by averaging the obtained results.

As a preliminary test we verify the intuition by which even advanced human presence detectors, taken alone, might fail in correctly classifying Over-the-Shoulder shots. Classification performance of OoS shots using a feature vector combining the pedestrian, the upper-body, and the face detection features only are shown in Table II for both movies, where results are presented in terms of precision, recall, and F1-measure. As a matter of fact, whenever the detectors are able to spot the presence of both characters, the classification precision becomes high. However this happens not very often (low recall) mostly due a missed detection of character B (the person filmed from behind), which is undoubtedly the most characterising OoS feature, and the most difficult to automatically capture.

The second test demonstrated the determinant role played by the set of saliency features in severely improving detection
performance of such difficult shot type. Classification performance of Over-the-Shoulder shots employing all features are shown in Table III for both movies, where results are presented in terms of precision, recall, and F1-measure.

<table>
<thead>
<tr>
<th>Movie title</th>
<th>F1 (%)</th>
<th>Prec (%)</th>
<th>Rec (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L'avventura (1960)</td>
<td>65.51</td>
<td>73.60</td>
<td>64.82</td>
</tr>
<tr>
<td>Il deserto rosso (1964)</td>
<td>68.90</td>
<td>72.26</td>
<td>71.02</td>
</tr>
</tbody>
</table>

Results show high ability in locating OtS shots by employing the combined set of features, obtaining an average precision superior to 70% for both movies, while both F1-measures remain above 65%, which outperforms the results declared in the only similar study on OtS in [10], where F-measure on OtS was around 0.51%, even if obtained on different material (which is not publicly available for direct comparison). Table III reports classification results only on OtS since performance on non-OtS shots, due to the numerosity of the class, are beyond 99% on all measures (precision, recall, F1-measure).

The last test employs all frames taken from L'avventura (resp. Il deserto rosso) as training set and considers Il deserto rosso (resp. L'avventura) for testing. In this case obtained performance in terms of F1-measure are 66.72% (resp. 66.67%) despite a severe drop of 23.60% (resp. 22.26%) in precision values, mostly due to the very diverse nature of film material (colour vs. grayscale). In ultimate analysis, the proposed technique demonstrates to work well on a large test set - almost 15000 tested frames - on the rich variety of scene composition and situations offered by art films, on both “black-and-white” and colour movie material, and with very few training samples.

V. CONCLUSIONS

In this paper we proposed a method for Over-the-Shoulder shot detection. The automatic recognition of this particular shot category, usually employed by directors to raise the dramatic tension of the narrative, is relevant since previous studies on cinema suggest that peculiar usage of shot types may be an important identifier of an individual director. In order to validate this hypothesis it is necessary to produce automatic recognition of all shot scales on large movie corpora, thus including the analysis of OtS, which is probably the most difficult shot scale to capture. The particular nature of this angle, shooting a character with the camera filming over the shoulder of another one, makes OtS recognition difficult even employing a series of state-of-the-art human presence detectors, such as face, upper-body, and pedestrian classifiers. To enhance classification performance we proposed a set of saliency features based on colour and motion, and combined them with human presence detectors in a SVM learning scheme. Despite the small number of OtS frames (~3%) in the dataset, overall OtS classification F1-measure is around 70%, so that the system proved to be quite robust to the scarcity of the training set, and well adapting to different experimental situations. The method has wide applicability, since it similarly performs on both colour or “black-and-white” movies, and well behave even if applied to art films, which are known as one of the most challenging material for automatic movie analysis, due to the variety of experimental aesthetic situations.

REFERENCES