Editorial Manager(tm) for Multimedia Tools and Applications
Manuscript Draft

Manuscript Number: MTAP1728R1

Title: Classifying Cinematographic Shot Types

Article Type: *CBMI'2011

Keywords: Shot type; camera distance; feature extraction; movie analysis.

Corresponding Author: Luca Canini

Corresponding Author's Institution:

First Author: Luca Canini

Order of Authors: Luca Canini;Sergio Benini;Riccardo Leonardi

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.

Manuscript - Answers to reviewers
"Classifying Cinematografic Shot Types"

Luca Canini, Sergio Benini, and Riccardo Leonardi

I. Introduction

First of all, we would like to thank the reviewers for their useful and detailed comments. On the basis of their remarks, and despite the short two-week period assigned for implementing all the suggested changes, the overall paper quality has been improved and the work organisation rearranged so as to improve the clarity of presentation and the scientific value of the results. Answers to all reviewers' requests follow.

II. Reviewer #1

A. Answers to Major Requests

1) *"This work represents a slight improvement of the system proposed by the authors in ICME 2010, including in this case the spectral amplitude feature as well as the classification with decision trees. However, according to the results, the decision trees do not perform better than the SVMs and the inclusion of such classifier should be better motivated."*

   Classification experiments involving C4.5 decision trees are here presented as complementary tests to those performed with SVM, so that the combination of the two allows for a more complete view on the effectiveness of the proposed approach. In general, even if SVM ensures higher classification accuracy, its main downside is found in the difficulties of parameter handling and model comprehension by the user. For example with SVM it is not easy to highlight the relevance of single features, which is usually important for problem understanding. Conversely, C4.5 builds a decision tree with a very intuitive procedure: each node of the tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Most important, at each node the feature that best divides the training data is chosen, thus pointing out the relevance of single features and their inter-relationships. The presented experiments comparing the effectiveness of different feature descriptors as input for classifiers are therefore crucial for future improvements of this approach and its integration in a possible longer toolchain towards semantic analysis of fiction content. A motivated explanation for the insertion of C4.5 decision tree has been inserted in the related section of the paper.

2) *"From my point of view the paper would benefit on a more detailed discussion about the possible utility and applications of the proposed approach."*

The shot type classifier can be exploited for many applications. As an example, on the basis of the proposed shot type classifier, the work in [1] 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 [11], 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 [9] 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 [2]) able to combine shots according to filmic grammar rules could undoubtedly benefit of such annotated content.

Hints to the utility and future applications of the proposed approach have been added to the last section of the revised paper.

3) *"It seems that the results could depend on the correct detection of the faces in the image. How many of the selected shots contained faces and how many of them did not? The most common implementation of the face detection algorithm applied [3] usually works well for frontal faces but not in other cases. Is that the case in this system? How does this affect the system given the selected shots?"*

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.

However, 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 for many reasons. As an example, Long 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 by state-of-the-art face detectors. Therefore we would tend to state that no human faces are actually present (read "detectable") in our database of Long shots. In addition to this, one of the considered movies, "Home" by Yann Arthus-Bertrand contains a large majority of people-less shots, and most of them are of the Long type. 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%.

Regarding the implementation of the Viola-Jones' algorithm, and the fact that it usually works well for frontal faces only, the last implementation on the OpenCV libraries [4] also comes with several cascade files for detecting profile faces, even if with slightly lower performance. Though accepting that failing the profile face detection might affect the system performance, we believe that the proposed multi-feature approach will tend to masquerade the weakness of a single feature, and that other features will anyway help in correctly assessing the shot type. Because of that, we tend to believe that a further numerical evaluation only of the non-frontal missed faces in our database would go beyond the scope of this work and would be of limited interest for the reader.

Details on face presence in the shot database have been added in the related paper sections.

4) *"It is unclear for me how if the trained classifiers are a set of individual binary classifiers for the three types of shots or multiclass classifiers. In the first case, how many positive and negative samples are employed for each classifier? If we have 1/3 LS shots, 1/3 MS shots and 1/3 CU shots we are dealing, for each category, with 1/3 positive and 2/3 negative samples which could bias the obtained results. Moreover I would prefer to see the precision/recall results. If dealing with multiclass classifiers I think a confusion matrix per classifier would better depict the results."*

SVM methods are binary. Thus in the case of multi-class problems, one must reduce the problem to a set of multiple binary classifications [5]. As a consequence two main strategies are possible: building binary classifiers which distinguish between one of the labels to the rest (one-against-all), or between every pair of classes (one-against-one). In our work we adopted the one-against-one approach; in this case, classification is done by a max-wins voting strategy, in which every classifier assigns the instance to one of the two classes, then the score for the assigned class is increased by one vote, and finally the class with the highest score determines the instance classification. Nevertheless when dealing with more than two classes, as in our case, people usually refer to multiclass SVM as a "single entity". SVM rely on support vectors, which are generally a small amount of the training data playing a fundamental role in the determination of the separation hyperplane(s). This procedure makes SVM relatively robust against unbalanced datasets.

The corresponding confusion matrices are reported in Table I for both classifiers (SVM and C4.5) and both image categories (man-made and natural, respectively). Details about Precision and Recall (which are already in the paper in the aggregated form of F-measure) are instead given in Table II.

TABLE I

Confusion matrices for SVM and C4.5 classifiers.

SVM-Man	**LS**	**MS**	**CU**
LS (gt)	0.691	0.198	0.110
MS (gt)	0.192	0.628	0.180
CU (gt)	0.081	0.070	0.849

SVM-Nat	**LS**	**MS**	**CU**
LS (gt)	0.951	0.023	0.027
MS (gt)	0.447	0.461	0.092
CU (gt)	0.384	0.040	0.576

C4.5-Man	**LS**	**MS**	**CU**
LS (gt)	0.663	0.194	0.143
MS (gt)	0.251	0.524	0.224
CU (gt)	0.109	0.106	0.785

C4.5-Nat	**LS**	**MS**	**CU**
LS (gt)	0.818	0.106	0.076
MS (gt)	0.342	0.316	0.342
CU (gt)	0.365	0.095	0.540

Corrections have been added in the related section of the paper, so as to increase the clarity of the explanation. Moreover, confusion matrices as requested by the reviewer have been inserted in the experimental section. Finally precision and recall figures have been also added in the paper, despite they are usually more appropriate for assessing performance of retrieval systems than classification problems as in this case.

TABLE II

Precision, Recall, Accuracy for SVM and C4.5 classifiers.

SVM-Man	**Precision**	**Recall**	**Accuracy**
LS	0.653	0.691	0.828
MS	0.695	0.628	0.808
CU	0.824	0.849	0.851

SVM-Nat	**Precision**	**Recall**	**Accuracy**
LS	0.754	0.951	0.796
MS	0.761	0.461	0.888
CU	0.837	0.576	0.856

C4.5-Man	**Precision**	**Recall**	**Accuracy**
LS	0.579	0.664	0.791
MS	0.616	0.524	0.762
CU	0.774	0.785	0.801

C4.5-Nat	**Precision**	**Recall**	**Accuracy**
LS	0.750	0.818	0.744
MS	0.375	0.316	0.803
CU	0.603	0.547	0.778

B. Answers to Minor Requests

1) *“Page 6, line 16: “techniques able to directly estimate the shot type starting from single images are not numerous”. What about works not dealing with single images? Page 7, “the required process of image segmentation and object recognition is often too computationally expensive”. It would depend on the requirements of the application. It would be interesting to anlyze those more complex systems as well in the SoA.”*

The problem of depth estimation from multiple images has been intensively investigated in the computer-vision research community [7] and the related literature is vast. In general multi-view approaches estimate disparity images for a robust depth estimation adopting different techniques, often using rectified views of the original images. Interesting approaches based on similar techniques are recently proposed to create depth maps for 3D-TV systems, such as in [6] and [8].

However, authors remain a bit sceptical about inserting this state-of-the-art on multi-view depth estimation in the new paper version, since when compared to stereo vision approaches, the here proposed method solves a very different problem: qualitatively classifying a shot instead of quantitatively estimating depth for each pixel. Thus, such comparisons with the related multi-view literature does not make much sense and could mislead the readers. For the sake of the truth the very first version of the ICME 2010 conference paper contained a few references to stereo approaches to the problem of depth estimation, and were (correctly, in our a-posteriori opinion) strongly criticised by the reviewers who suggested their removal. Concluding, we apologise for this position, but before inserting this part of SoA in the paper we appeal to the Associate Editor's opinion.

2) *"Section 3: Local Distribution of Color Intensity: It seems that this descriptor will be quite dependant on the resolution of the original video, this should be taken into account, specially with fixed parameters such as the dimensions of the sliding window.*

All movies used for experiments have been pre-processed in order to have all videos at the same starting resolution of $\mathcal{W} \times \mathcal{H}$, with $\mathcal{W} = 720$ and $\mathcal{H} = 480$. The sliding window w_I of dimensions $\frac{\mathcal{W}}{R} \times \frac{\mathcal{H}}{R}$ (where $\mathcal{R} = 20$) has been chosen so as to maximise classification performance (evaluated with the SVM classifier on the single feature for different values of $\mathcal{R}$), but as stated in the paper, $\mathcal{R}$ is tuneable.

3) *"Page 9, lines 42-53: From my point of view the assumptions made in those lines are arguable, for example "scene points farther with respect to the camera will be the darkest ones"*

This and similar expressions have been softened through the whole paper.

4) *"Section 4: motion activity maps: I assume that authors are working with motion vectors extracted from the coded stream. What codec are they using? How the coding parameters could influence this measure? What about cases where camera motion is present instead of moving object with static camera? Have been this kind of situations included in the content set?"*

Motion vectors are extracted from predicted frames of the MPEG-4 compressed stream. After a suitable filtering process (a texture filter followed by a median filter) to remove motion vectors that tend to be errant due to the typical noise of block-based motion vector estimation, motion activity maps are computed. 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. Again, the idea is having another simple descriptor, that might lead to rough classification score if taken alone, but that when combined with others, contributes to discriminate the shot type in most situations.

This clarification has been now inserted also in the revised version of the paper.

5) *"Authors mention a preclassification step between natural and man-made shots and the results are presented individually for both categories. Are those results obtained assuming a perfect classification between both categories? It is not clear if the final results consider the possible missclassification of the original shots. [from another question, here merged] With respect to the division between natural and man-made shots. Why a pre-processing step is introduced? I think that the classification procedures (SVMs or decision trees) should be able to deal with such distintion if they were provided with the spectral information. Details about the training and classification procedure of this pre-processing step are missing."*

In the paper we mention a pre-classification step between natural and man-made shots, which can be performed automatically by using the method proposed by Torralba et. al. in [10], which makes use of the spectral feature as a discriminant factor. Details about the training and classification procedures of this pre-processing step can then be found in the same published work, 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. A clarifying sentence has been added in the related paper section.

It is indeed true that SVMs or decision trees should be able to deal with the distinction between man-made and natural if they were provided with the spectral information. However, apart from the description in the aforementioned work, the beneficial step of preliminary subdividing the two image categories has been also confirmed by an attempt we made to have a one-step classification without distinguishing between man-made and natural images, which in fact returned lower classification performance on the shot type.

III. Reviewer #2

A. Answers to Requests

1) *"The introduction frames the film characteristic subjected to study (shot type) within a few more features part of the cinematographic grammar. However these are part of a greater taxonomy of movie stylistic capabilities; for the sake of completeness some mention of the whole scheme, including additional visual features (color, framing, lightning, composition) as well as sound and higher-order entities (rhythm, editing, continuity/discontinuity), etc. Some bits are mentioned from time to time (such as the references to establishing shots), but a simple enumeration in the introduction would help to better establish the context of the study."*

   In the new revised version of the paper the introduction has been enriched with other elements referring to the taxonomy of movie stylistics. Since the aspects covered by this topic are numerous, and one paper introduction clearly cannot span over its totality, we add here (and in the paper) a few more bibliographic references in addition to the Arijon's work in [11], such as [12] [13] [14] [15] and [16], that can be helpful for the interested reader.

2) *"In p.6, l.20 says that the method does not need to compare different images of the same scene. There is an exception: for the motion descriptors, images do need to be compared (to extract the motion between them)."*

   Motion vectors are in fact extracted from predicted frames of the MPEG-4 compressed stream, so there is no need to compare different images from the same scene (the process is completely transparent for authors). The sentence actually refers to the fact that the proposed technique works on monocular images rather than relying on multi-view images of the same scene.

The paper has been modified in the related section so as to improve clarity with respect to this request.

3) *"In p.9, l.8, it is mentioned that objects near to the camera are sharper. What about the camera focusing on distant objects with near objects unfocused? It is a well known stylistic resource."*

   Authors agree that in the mentioned scenario the proposed feature contribution to the shot type classification would be erroneous. In this case other descriptors, such as the motion activity maps, might help to discriminate the correct shot type anyway. It is also true that, according to our experience, directors bring into play this stylistic resource not so often to justify the building of an ad-hoc feature or mechanism to detect this particular situation.

4) *"I appreciate the well-chosen examples in Fig. 2, in which it can be seen that also long shots may contain some important sections in the high variance area."*

   Thanks about that!

5) *"In p.13, l.40 I disagree with the statement: a Mam, as described, does not really show variation distribution across the shot, since it computes an average. It would produce the same value for a short high activity peak and then still frames than for a shot containing continuous and stable low movement. So it shows average activity, but not variations of motion."*

   Authors fully agree with the comment. The sentence has been rephrased in the paper as: "[...] a Mam expresses the behaviour of motion activity in the shot by displaying its *average* temporal distribution over the shot duration.".

6) *"In p.15, l.50: shouldn't this measurement be also a function of camera angle, in addition to camera distance? That is, on frontal-looking camera it works as intended, but if the camera is, say tilted, it would add additional distortion."*

The reviewer spotted an interesting case that was not initially included in the paper, but that is actually under development for the new version of the classifier. In the paper, we actually claim that we do not “tackle a precise detection of vanishing points”, but “we rather aim at finding an estimator of camera distance by collecting slopes of perspective lines in shot key-frames.” At the moment of writing this review, we are trying to perform a check on the position of vanishing point(s) to verify that it (they) is (are) located between the frame top and bottom lines, i.e. that the camera is frontal-looking or, let’s say, tilted within a small angle interval. Were this conditions not verified, we would not take the related descriptor into account during the classification step. According to our preliminary experimental evidences, however, the situation of having large angles of camera tilt in filmic material is quite rare.

7) *“The dataset (p. 19): only the number of shots are mentioned. How many movies are used? What is the shot amount taken from each movie? How are shots sampled from each movie? (randomly, first N shots, first N shots that satisfy the LS/MS/CU classification, etc) What cinematographic genres do the sampled movies span?”*

Our database is composed by 12 movies spanning the main genres of modern cinematography. They are listed in the following along with their genres as defined by IMDb [18]:

- Indiana Jones and the Last Crusade - (Action/Adventure)
- War of the Worlds - (Action/Adventure/Drama)
- A Beautiful Mind - (Biography/Drama)
- All or Nothing - (Drama/Comedy)
- Home - (Documentary)
- Eternal Sunshine of the Spotless Mind - (Drama/Romance/Sci-Fi)
- Spring, Summer, Fall, Winter... and Spring - (Drama)
- Samaritan Girl - (Drama)
- Phone Booth - (Mystery/Thriller)
- Seven Swords - (Action/Fantasy)
- Once Upon a Time in the West - (Western)
- All about my mother - (Drama)

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 a frame form the data and we classify it as Long, Medium, Close-up (or "not relevant"). The process ends when a database with 1000 samples for each of the three classes: Long, Medium and Close-up has been gathered.

These details have been inserted also in the related section of the paper.

8) *"P.20, l.44: please confirm the source of the performance gap by giving the error rates for the 3 shot types. Also, the mentioned large share of close-ups in the database would be 33%, as explained before. Is it a larger share than the usual amounts in movies? In other words, which are the prior probabilities of LS/MS/CU in standard movies? (this is probably genre-dependent, and can vary significantly from one movie to another; still, some rough estimation perhaps on the same 25 IMDb movies used before for the artificial/natural split would help to frame the detection problem)."*

The requested correct rates in natural/man-made classification for the three shot types are 74% (LS), 85% (MS), and 90% (CU), respectively. The same figures have been also inserted in the related section in the paper.

For what concerns the prior probabilities of LS/MS/CU in standard movies, the reviewer correctly points out that these percentages vary significantly from one movie to another and from genre to genre. An automatic estimation on the complete movie database 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 [1] where we analyse 83 "great movie scenes" chosen to represent popular films from 1958 to 2009 (total duration of more than 3 hours of video and 2311 shots).

9) *"P. 21, l. 10: could you give a brief explanation of the "stratification process"?"*

In a common classification scenario data are divided into two sets: one used for training and the other used for testing the model. If the subdivision is performed by a completely random approach, the training or the test set may not be representative for the overall data set. A very extreme example is the following: let data be split into two classes, 100 elements for each class. If 120 elements are used for training and 80 for testing, divided by a random algorithm, it may be that the training set contains only 20 elements of one class, thus producing a model polarized towards the other class (if 20 are not enough to represent the variability of a class). To avoid this problem, one should take care of the fact that each class should be correctly represented in both the training and testing sets. This process is called stratification. Therefore stratification is the process of rearranging the data as to ensure each fold is a good representative of the whole. In the mentioned example the training set would contain 60 element for each class, while the test set 40.

10) *"P. 22: in Table 1 natural LS seems to be the worst performing case. It would be nice to show the confusion matrix, this would help in finding out which are the largest misclassifications made (i.e. which wrong assignments are the most frequent)."*

See Answer no. 4 given to Reviewer #1, where the requested confusion matrix is provided.

11) *"P. 24, l. 20 please explain the training process. What kind of cross validation is done? Are both datasets combined? Is the test set included into training at all?"*

The two datasets are not combined in any manner, i.e. there are two SVM multiclass classifiers, one for Natural and the other for Man-made images. Moreover, the two employed test sets are not included at all in the two distinguished training processes (one performed for Natural and the other for Man-made images, respectively).

Once clarified the previous point, we would like to spend a word on the cross-validation process. SVM allows for two completely different (in their aims) cross-validations:

a) The first one, and more general, (i.e. it can be applied to almost every classifier) is performed during the classification stage: the training and test sets are crossed-over in successive rounds such that each data point has a chance of being validated against. This basic form of cross-validation is called k-fold cross-validation. In the specific, data are first partitioned into k equally (or nearly equally) sized segments or folds. Subsequently

k iterations of training and testing are performed such that within each iteration a different fold of the data is held-out for testing while the remaining $(k-1)$ folds are used for training. This method is usually employed when data may not be enough for performing a simple subdivision in training and test set. This is not our case, since we are confident that given the chosen machine learning algorithms, our data is enough. So in our paper we do not refer to this method.

b) The second scenario in which cross-validation is applied is typical of SVM and considers the training set only. Its goal is to select the best possible parameters for the trained model. This is achieved by performing the method described in a) (notice: only on the training set!) for each possible combination of parameters, in order to have a heuristic assessment of what the performance of the SVM could be on a real test set given the chosen parameters. Once the best set of parameters has been selected by applying cross-validation (again, on the training set only), the whole training set is used for the learning phase while the evaluation is performed on the (unused, until now) test set. In our paper, we do refer to this kind of cross-validation (see [19] fur further reading).

12) *"P. 24, l. 52 The statement holds for all results except for specificity in natural LS, which is moreover significantly lower than for all the other classes. Would it be possible to adventure an explanation for this fact?"*

A tentative explanation for the lower performance on LS shot on natural images is here provided. In general, with respect to all other shot types (MS and CU), Long shots have, by definition, no filmed main subject in the foreground on whose basis to compute the shot type. However, in the case of Long shots depicting man-made scenes, the background often presents structured subjects (such as buildings, etc.) which allow anyway for a correct classification of the camera distance category. This interpretation is partially confirmed by the high classification performance of the spectral feature on this shot type, as presented in Table 3. Conversely, Long shots of natural scenes provide neither foreground subjects nor background geometrical elements on whose basis to drive a decision on the shot type.
This comment has been also inserted in the related section of the paper.

13) *“P.25, l 48 Though the statement seems reasonable, would it be possible to show ground truth data from the the dataset of face presence by shot class?”*

See Answer no. 3 given to Reviewer #1.

14) *“P. 26, l.44 why is this feature the only split between artificial/natural for analysis? If it is because the others show no significant difference in performance across the two classes, it will be good to state it explicitly.”*

See Answer no. 5 given to Reviewer #1 (minor requests).

15) *“P. 27, l. 10 There is a significant difference in performance between the composite SVM and each individual SVM. However it is not clear which features are contributing most to the composite (maybe the best combination is not with the best performing individual ones but with the ones most complementary). Therefore some type of meta analysis comparing subsets of features would be great.”*

Even if SVM ensures higher classification accuracy, its main downside is found in the difficulties of parameter handling and model comprehension by the user. For example, as pointed out by the reviewer, with SVM it is not easy to highlight the relevance of single features, which is usually important for problem understanding.

When the feature set is numerous, one common approach to overcome this problem is to add a step which performs “feature selection” before the SVM classifier, that is to select a subset of relevant features for building a robust learning model. Feature selection also helps to acquire better understanding about data by revealing which are the important features and how they are related with each other (see for example [17] for a feature selection method based on information theory techniques).

In our case, being the problem very specific, we adopted an alternative approach: instead of developing a number of general purpose features and select few of them to feed the SVM, we tried to develop one reduced set of ad-hoc features which takes into account the different aspects and the specific nature of the given problem. The meta analysis suggested by the reviewer is somehow provided by the C4.5 decision tree analysis: the use of these classifiers is in fact meant as complementary to the experiments performed with SVM, so that the combination of the two allows for a more complete view on the effectiveness of the proposed approach. Differently from SVM, C4.5 builds a decision tree with a very intuitive procedure: each node of the tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Most important, at each node the feature that best divides the training data is chosen, thus pointing out the relevance of single features, their best order of application, and their complementary relationships. This analysis for example revealed the dominant role of single descriptors based on human faces A_F and the geometric composition of the scene α and the complementary (though still important) nature of the other proposed features.

16) *"Some typos/suggested grammatical improvements in the text:*
p.3, l.42 "intended as shot duration" -> "with the meaning of shot duration"
p.5, l.36 "the occupancy of space" -> "the space occupied" (or filled)
p.7, l.43 "rely on a limited number of sources of information" -> "rely on the availability of special sources of information" would probably be more appropriate
p.8, l.36 "they difficultly bring some information" -> awkward, maybe "they barely bring any information"
p.8, l.55 "build taxonomy" -> "build a taxonomy"
p.9, l.12 "no classification performance are" -> "no classification performance is"
p.9, l.30 "on images, allows for" -> "on images allows for" (remove comma)
p.25, l.6 "continuos" -> "continuous"
p.25, l.34: "in the specific, results show" -> "Specifically, results show"."

Almost the totality of these suggestions have been implemented in the new version of the paper. Thanks about that.

IV. Reviewer #3

A. Answers to Requests

1) *"To improve the quality of the paper there is need of better describing the types of shots that are part of the dataset: the definitions (page 18) imply a strong dependence on the presence of humans, so that one of the features used (presence of faces and face size) seems to have a strong impact because of that."*

   See Answer no. 3 to Reviewer #1.

2) *"The paper would benefit also from testing on some standard datasets like TRECVid or PASCAL VOC (perhaps not using the motion feature)."*

   The authors are seriously considering to broaden the experimental dataset with other video and movie repositories. However, a severe problem when dealing with this kind of data is that most of this material is covered with copyright, and no datasets are made easily available: in fact even if the purpose is to use data in ways that are eligible for fair use consideration, using copyrighted material could anyway turn into problematic issues.
   Unfortunately, the databases suggested by the reviewer do not contain filmic material. As far as we know, Pascal Voc provides standardised databases for object recognition in 20 different classes: person, bird, cat, cow, dog, horse, sheep, aeroplane, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining table, potted plant, sofa, tv/monitor. For what concerns Trecvid instead, various types of data have been involved in the last years, ranging from BBC rushes, to London Gatwick surveillance video files, broadcast news and huge quantities of internet material, but no considerable filmic material to perform experiments have been included yet.

References

[1] L. Canini, S. Benini, and R. Leonardi, "Affective Analysis on Patterns of Shot Types in Movies," in Proceedings of 7th International Symposium on Image and Signal Processing and Analysis (ISPA 2011), Dubrovnik, Croatia, September 4-6, 2011.

[2] L. Canini, S. Benini, and R. Leonardi, "Interactive Video Mashup Based on Emotional Identity," in Proceedings of the 2010 European Signal Processing Conference (EUSIPCO '10), Aalborg, Denmark, August 23-27, 2010.

[3] Viola, P., Jones, M. “Rapid object detection using a boosted cascade of simple features,” Proc. of CVPR (2001).

[4] Bradski, G., “The OpenCV Library” Dr. Dobb’s Journal of Software Tools, 2000.

[5] Hsu, C.-W. and Lin, C.-J., “A comparison of methods for multi-class support vector machines,” IEEE Transactions on Neural Networks, 13(2):415–425, 2002.

[6] P. Kauff, N. Atzpaadin, C. Fehn, M. Mueller, O. Schreer, A. Smolic, and R. Tanger. Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability. Signal Processing: Image Communication, 22(2):217–234, 2007.

[7] D. Scharstein, R. Szeliski, and R. Zabih, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” In IEEE Workshop on Stereo and Multi-Baseline Vision, pages 131–140, Dec 2001.

[8] S. Yea and A. Vetro. CE3: Study on depth issues. ISO/IEC JTC1/SC29/WG11 and ITU SG16 Q.6 JVT-X073, 2007.

[9] J. Porteous, S. Benini, L. Canini, F. Charles, M. Cavazza and R. Leonardi, ”Interactive Storytelling Through Video Content Recombination,” in Proceedings of ACM Conference on Multimedia (ACM MM ’10), Florence, Italy, October 25-29, 2010.

[10] Torralba, A., Oliva, A.: Depth estimation from image structure. IEEE Trans. on PAMI 24(9), 1226–38 (2002).

[11] Arijon, Daniel (1976): Grammar of the Film Language. London: Focal Press

[12] Bordwell, D. and Thompson, K. (1993): Film Art: An Introduction. New York: McGraw Hill

[13] Izod, J. (1984): Reading the Screen (York Handbooks). Harlow: Longman

[14] Millerson, G. (1985): The Technique of Television Production. London: Focal Press

[15] Monaco, J. (1981): How to Read a Film. New York: Oxford University Press

[16] Sobchack, T. and Sobchack, V. C. (1980): An Introduction to Film. Boston: Little, Brown and Company

[17] Peng, H.C., Long, F., and Ding, C., Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, pp. 1226–1238, 2005.

[18] Internet Movie Database, *http://www.imdb.com/*

[19] Refaeilzadeh, P. and Tang, L. and Liu, H., Cross Validation, In Encyclopedia of Database Systems, 2009.

Noname manuscript No.
(will be inserted by the editor)

Classifying Cinematographic Shot Types

Luca Canini · Sergio Benini · Riccardo Leonardi

Received: date / Accepted: date

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

L. Canini
Department of Information Engineering, University of Brescia
Tel.: +39 030-371-5511
Fax: +39 030-380014
E-mail: luca.canini@ing.unibs.it

S. Benini
Department of Information Engineering, University of Brescia
Tel.: +39 030-371-5528
Fax: +39 030-380014
E-mail: sergio.benini@ing.unibs.it

R. Leonardi
Department of Information Engineering, University of Brescia
Tel.: +39 030-371-5534
Fax: +39 030-380014
E-mail: riccardo.leonardi@ing.unibs.it

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.

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 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 [2].

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).

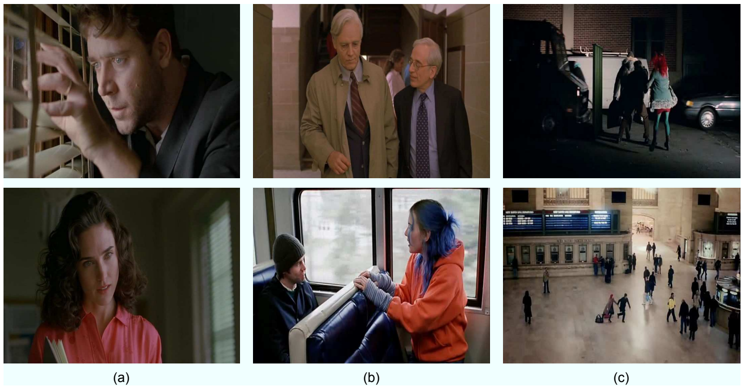

Fig. 1 Shot types: a) Close-ups, b) Medium and c) Long shots, as in [2].

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 Figure 1-a, depicting human characters from the breast upwards.

In a Medium shot, as in the case of the standing actors depicted in the examples of Figure 1-b, 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 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 Figure 1-c) 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 [2] 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 [2], [6], 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) [13]. 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 [17]. 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, 4, 5, 6 and 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 [14] and in [16].

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 [2]. 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 [8], texture gradients [35], edges and junctions [3], 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 [18], 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 [4] only other two works in literature directly dealt with shot type detection in movies. The work in [12] 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 are 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 [11], 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 [11] 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 $\frac{\mathcal{W}}{\mathcal{R}} \times \frac{\mathcal{H}}{\mathcal{R}}$ (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 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 Figure 2 for a Long shot in (a) and a Close-up in (b).

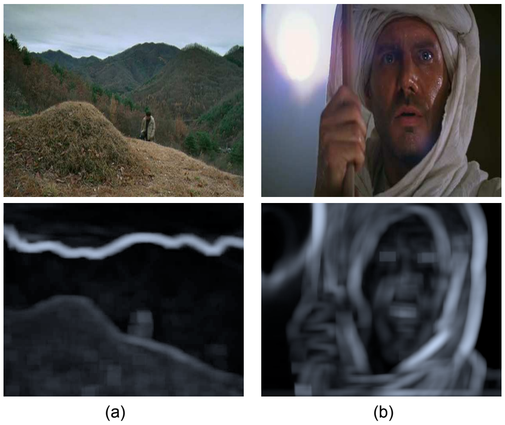

Fig. 2 Examples of original images and the obtained histogram variance images for a) a Long shot and b) a Close-up.

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 Figure 2-a. 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 [5], 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:

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

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 Figure 3-a a key-frame extracted from the movie “Raiders of the Lost Ark” is given, together with a representation of the associated motion field. In Figure 3-b, 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.

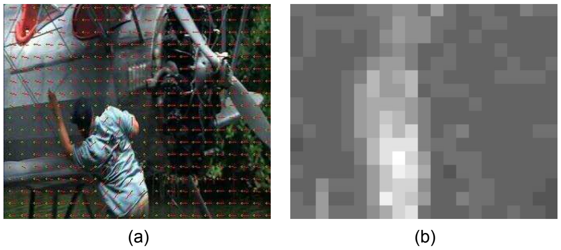

Fig. 3 a) The motion vector field of a key-frame from “Raiders of the Lost Ark” and b) its corresponding Mam.

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.

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 [15] 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

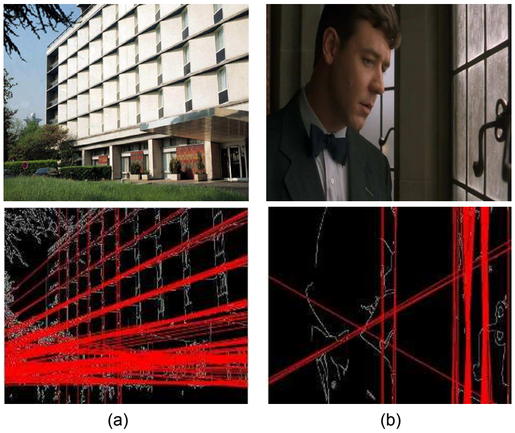

Fig. 4 Examples of perspective lines extracted by the Hough transform a) from a Long shot and b) from a Close-up.

point, as in Figure 4-a). Conversely, inclinations of perspective lines evidently differ when observing a less distant shot (Figure 4-b).

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.

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 [7] also comes with several cascade files for detecting profile faces, even if with slightly lower performance. In Figure 5 an example of the output provided by the Viola-Jones method is given, where detected faces are highlighted by bounding boxes.

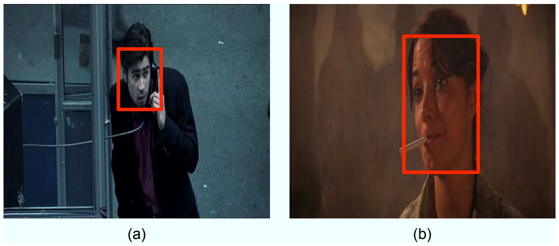

Fig. 5 Examples of faces (red bounding boxes) a) in a MS and b) in a CU.

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 Figure 5, (a) is classified as MS, while (b) is a CU.

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 Figure 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 (Figure 6-a), a man-made one has sharp dominant vertical and horizontal components due to the presence of geometrical artificial structures (Figure 6-b).

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 Equation 1

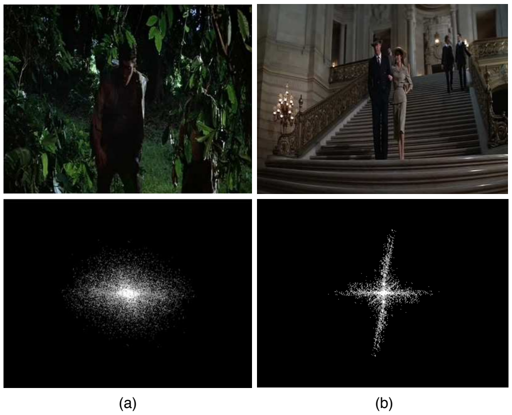

Fig. 6 Examples of global magnitude of the Fourier transform of a) a natural and b) a man-made image (the white plots represent the 80% of the energy).

as proposed in [27]:

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

where θ is the phase of the frequency vector $\overline{f}$, $\varLambda(D,\theta)$ is a magnitude factor, and $\varGamma(D,\theta)$ 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:

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

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) [13]. 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.

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) [1], 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

[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	*Indiana Jones and the Last Crusade*	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

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 1000 Long shots, 1000 Medium ones, and 1000 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 [10] where we analyse 83 “great movie scenes” chosen to represent popular films from 1958 to 2009 (total duration of more than 3 hours 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.

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 sub-sets. 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 performance 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*).

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.

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**

In both testing scenarios (natural and man-made) fair performance are 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

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

first three levels of the decision trees depicted in Figures 7 and 8 for man-made and natural images, respectively. From an inspection of both decision trees, it emerges the predominant

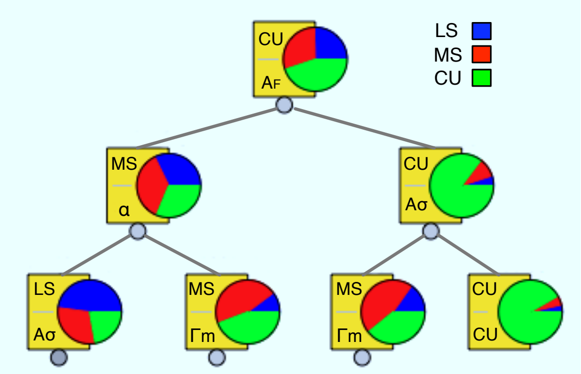

Fig. 7 The first three levels of the decision tree for man-made images. In each node the splitting feature is shown, together with the majority class, and a pie diagram with the distribution of different shot types.

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 Figure 7) the presence of faces is the most significant descriptor, they switch their positions in natural images (Figure 8) where the perspective aperture takes over. Other features intervene on deeper lev-

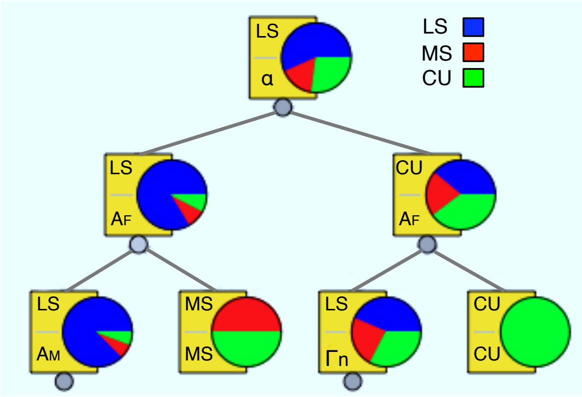

Fig. 8 The first three levels of the decision tree for natural images. In each node the splitting feature is shown, together with the majority class, and a pie diagram with the distribution of different shot types.

els 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 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 [19], 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 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, performance obtained on the testing datasets in terms of *accuracy*, *specificity*, *F-measure*, *precision* and *recall* are shown in Table 5.

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**

In both testing scenarios (natural and man-made) good performance are 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 general achieve a

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

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 general, with respect to all other shot types (MS and CU), Long shots have, by definition, no filmed main subject in the foreground on whose basis to compute the shot type. To be specific, in the case of Long shots depicting man-made scenes, the background often presents structured subjects (such as buildings, etc.) which allow anyway for a correct classification of the camera distance category. This interpretation will be partially confirmed by the high classification performance of the spectral feature on this shot type, as presented in Table 6 in the next section where we study the performance of single descriptors. Conversely, Long shots of natural scenes provide neither foreground subjects nor background geometrical elements on whose basis to drive a decision on the shot type, interpretation which is reinforced by the low specificity of the spectral feature on natural images which will be given in Table 6.

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

Table 6 Shot type classification results obtained with SVM using single features $\{A_\sigma\}$, $\{A_M\}$, $\{\alpha\}$, $\{A_F\}$, $\{\mathit{\Gamma}_m\}$ and $\{\mathit{\Gamma}_n\}$.

	A_σ (Colour)			A_M (Motion)		
Shot-type	**Ac.(%)**	Sp.(%)	F_1	**Ac.(%)**	Sp.(%)	F_1
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)		
	Ac.(%)	Sp.(%)	F_1	**Ac.(%)**	Sp.(%)	F_1
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
	$\mathit{\Gamma}_m$ (Spectral - Man-made)			$\mathit{\Gamma}_n$ (Spectral - Natural)		
	Ac.(%)	Sp.(%)	F_1	**Ac.(%)**	Sp.(%)	F_1
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

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 performance 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.

Since camera distance deeply affects the emotional involvement of the audience and the process of identification of viewers with the movie characters [2], 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 [10] 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 [2], 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 [9]) able to combine shots according to filmic grammar rules could undoubtedly benefit of such annotated shot content.

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

References

1. Internet Movie Database (IMDb). URL http://www.imdb.com/
2. Arijon, D.: Grammar of the Film Language. Silman-James Press (1991)
3. Barrow, H., Tenenbaum, J.: Interpreting line drawings as three-dimensional surfaces. Artificial Intelligence **17**(1-3), 75–116 (1981)
4. Benini, S., Canini, L., Leonardi, R.: Estimating cinematographic scene depth in movie shots. In: Proceedings of the IEEE International Conference on Multimedia & Expo (ICME). Singapore (2010)
5. Benini, S., Xu, L.Q., Leonardi, R.: Using lateral ranking for motion-based video shot retrieval and dynamic content characterization. In: Proc. of CBMI. Riga, Latvia (2005)
6. Bordwell, D., Thompson, K.: Film Art: An Introduction. McGraw-Hill (1997)
7. Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000)
8. Brooks, M.J.: Shape from shading. MIT Press, Cambridge, MA, USA (1989)
9. Canini, L., Benini, S., Leonardi, R.: Interactive video mashup based on emotional identity. In: Proceedings of the 2010 European Signal Processing Conference (EUSIPCO). Aalborg, Denmark (2010)

10. Canini, L., Benini, S., Leonardi, R.: 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 (2011)

11. Cantoni, V., Lombardi, L., Porta, M., Vallone, U.: Qualitative estimation of depth in monocular vision. In: Proc. of IWVF, pp. 135–144. Springer-Verlag, London, UK (2001)

12. Cherif, I., Solachidis, V., Pitas, I.: Shot type identification of movie content. In: Proceedings of International Symposium on Signal Processing and its Applications. Sharjah, United Arab Emirates (2007)

13. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3), 273–297 (1995)

14. Duan, L.Y., Xu, M., Yu, X.D., Tian, Q.: A unified framework for semantic shot classification in sport videos. In: Proc. of ACM MM, pp. 419–420. ACM, New York, NY, USA (2002)

15. Duda, R.O., Hart, P.E.: Use of the hough transformation to detect lines and curves in pictures. Commun. ACM **15**(1), 11–15 (1972)

16. Ekin, A., Tekalp, A.M.: Robust dominant color region detection and color-based applications for sports videos. In: Proc. of ICIP'03, pp. 1025–1028. Barcelona, Spain (2003)

17. Hanjalic, A.: Extracting moods from pictures and sounds. IEEE Signal Processing Magazine **23**(2), 90–100 (2006)

18. Hoiem, D.: Seeing the world behind the image: Spatial layout for 3d scene understanding. Ph.D. thesis, Robotics Institute, Carnegie Mellon Univ., Pittsburgh, PA (2007)

19. Hsu, C.W., Lin, C.J.: A comparison of methods for multi-class support vector machines. IEEE Transactions on Neural Networks **13**(2), 415–425 (2002)

20. Jeannin, S., Divakaran, A.: Mpeg-7 visual motion descriptors. IEEE Trans. on CSVT **11**(6), 720–724 (2001)

21. Keller, J.M., Crownover, R.M., Chen, R.Y.: Characteristics of natural scenes related to the fractal dimension. IEEE Trans. on PAMI **9**(5), 621–627 (1987)

22. Kotsiantis, S.B.: Supervised machine learning: A review of classification techniques. Informatica **31**, 149–268 (2007)

23. Kurita, T., Otsu, N., Abdelmalek, N.: Maximum likelihood thresholding based on population mixture models. Pattern Recognition **25**(10), 1231–1240 (1992)

24. Matessi, A., Lombardi, L.: Vanishing point detection in the hough transform space. In: Proc. of Euro-PAR '99, pp. 987–994. Springer-Verlag, London, UK (1999)

25. Monaco, J.: How To Read A Film. New York: Oxford University Press (1981)

26. Nagai, T., Naruse, T., Ikehara, M., Kurematsu, A.: Hmm-based surface reconstruction from single images. In: Proc. of ICIP'02, pp. 561–564. Rochester, NY, USA (2002)

27. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. International Journal of Computer Vision **42(3)**, 145–175 (2001)

28. Palmer, S.E.: Vision Science-Photons to Phenomenology. MIT Press, Cambridge, MA (1999)

29. Porteous, J., Benini, S., Canini, L., Charles, F., Cavazza, M., Leonardi, R.: Interactive storytelling via video content recombination. In: Proceedings of ACM Conference on Multimedia (ACM MM). Florence, Italy (2010)

30. Quinlan, J.R.: C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning), 1 edn. Morgan Kaufmann (1993)

31. Quinlan, J.R.: Improved use of continuous attributes in c4.5. Journal of Artificial Intelligence Research **4**, 77–90 (1996)

32. Refaeilzadeh, P., Tang, L., Liu, H.: Cross validation. In: In Encyclopedia of Database Systems (2009)

33. Salt, B.: Moving into Pictures. More on Film History, Style, and Analysis. Starword, London (2006)

34. Shimshoni, I., Lindenbaumlpr, M., Moses, Y.: Shape reconstruction of 3d bilaterally symmetric surfaces. Proc. of ICIAP **0**, 76 (1999)

35. Super, B.J., Bovik, A.C.: Shape from texture using local spectral moments. IEEE Trans. on PAMI **17**(4), 333–343 (1995)

36. Torralba, A., Oliva, A.: Depth estimation from image structure. IEEE Trans. on PAMI **24**(9), 1226–38 (2002)

37. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. Proc. of CVPR (2001)

38. Wang, H.L., Cheong, L.F.: Taxonomy of directing semantics for film shot classification. IEEE Transactions on Circuits and Systems for Video Technologies **19**, 1529–1542 (2009)

39. Xie, L., Chang, S.F., Divakaran, A., Sun, H.: Structure analysis of soccer video with hidden markov model. In: Proceedings of ICASSP'02. Orlando, Florida, USA (2002)

40. Zeng, W., Gao, W., Zhao, D.: Video indexing by motion activity maps. In: Proc. of ICIP. Rochester, USA (2002)

Luca Canini got 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 was born in Verona, Italy. He has 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.

Prof. 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.

***Author Photographs**

Click here to download high resolution image

***Author Photographs**

Click here to download high resolution image

***Author Photographs**

Click here to download high resolution image