USING CONTEXT SALIENCY FOR MOVIE SHOT CLASSIFICATION*

Min Xu[1,2], *Jinqiao Wang*[2], *Muhammad A. Hasan*[1]
Xiangjian He[1], *Changsheng Xu*[2], *Hanqing Lu* [2], *Jesse S. Jin*[3]

[1]Centre for Innovation in IT Services and Applications, University of Technology, Sydney, Australia
[2]Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[3]School of Design, Communication and I.T., University of Newcastle, NSW, Australia

ABSTRACT

Movie shot classification is vital but challenging task due to various movie genres, different movie shooting techniques and much more shot types than other video domain. Variety of shot types are used in movies in order to attract audiences attention and enhance their watching experience. In this paper, we introduce context saliency to measure visual attention distributed in keyframes for movie shot classification. Different from traditional saliency maps, context saliency map is generated by removing redundancy from contrast saliency and incorporating geometry constrains. Context saliency is later combined with color and texture features to generate feature vectors. Support Vector Machine (SVM) is used to classify keyframes into pre-defined shot classes. Different from the existing works of either performing in a certain movie genre or classifying movie shot into limited directing semantic classes, the proposed method has three unique features: 1) context saliency significantly improves movie shot classification; 2) our method works for all movie genres; 3) our method deals with the most common types of video shots in movies. The experimental results indicate that the proposed method is effective and efficient for movie shot classification.

Index Terms— Supervised learning, Support Vector Machines, Image classification, Feature extraction

1. INTRODUCTION

A video shot, which is created of a series of frames with high similarity, is normally used as a unit for video analysis, indexing and retrieval. Various video shot types are applied by video producer to convey video stories and attract audience attention. Therefore, video shot classification become very important for video content understanding and accessing. Most of the existing works of shot classification were performed in sports video domain [1, 2, 3, 4, 5] and news video domain [6]. Sports play field and motion features were used for sports video shot classification in [1, 4, 5]. Limited shot types, such as close-up shot, medium shot, full court shot and out of the field were usually classified for sports videos [2, 5, 7].

Recently, with the growth in production of movies and the increasing demand for personalized movie browsing, movie shot classification starts attracting research efforts [8, 9, 10]. Compared to sports videos and news videos, shot classification for movies is a challenging task based on two main reasons: 1) movie has much more shot types than sports video and news video; 2) sport video has very organized video structure and shot transition patterns due to the constrains from game rules. Movie video structure is unorganized. Various movie genres and different movie shooting techniques make it difficult to seek direct help from cinema knowledge.

Recent works either focus on a particular movie genre, e.g., action movie [10], or classify movie shot into directing semantic classes, such as zoom in, zoom out and panning, by motion features [8, 9]. In this paper, we propose to combine context saliency and color/texture features for movie shot classification. Different from the existing work, the proposed method has three unique features: 1) our method works for all movie genres; 2) our method deals with the most common types of video shots in movies [11]; 3) context saliency is used to significantly improve movie shot classification.

Visual attention has been proved meaningful for image/video analysis [12], especially for video highlight detection [13, 14] and image summarization [15]. In a movie, by using different shot types, the director intends to bring audiences various watching experience by attracting their attentions to different locations on the movie screen. Therefore, we propose to use the location and amount of visual attentions distributed on a movie screen with color and texture information to infer movie shot type.

The proposed method is briefly introduced in Section 2. Movie shot classes are introduced in Section 3. Section 4 proposes context saliency map generation for keyframes extracted from video shot. SVM based shot classification is presented in Section 5. Experiments are in Section 6. Finally, we conclude the paper in Section 7.

*THIS RESEARCH WAS SUPPORTED BY NATIONAL NATURAL SCIENCE FOUNDATION OF CHINA NO. 61003161, NO. 60833006 AND NO. 60905008, AND UTS ECR GRANT

2. OUR PROPOSED FRAMEWROK

The proposed method includes two main parts, i.e., saliency map generation and shot classification as shown in Fig. 1. Firstly, a contrast saliency map is calculated for each keyframe. In order to accurately measure the visual attention, we reduce redundant salient regions in contrast saliency map by calculating information density and incorporate geometric information to generate context saliency. Later, features related to saliency distribution are extracted from the generated context saliency map and then combined with gray histogram, texture entropy and correlation for SVM to achieve classification.

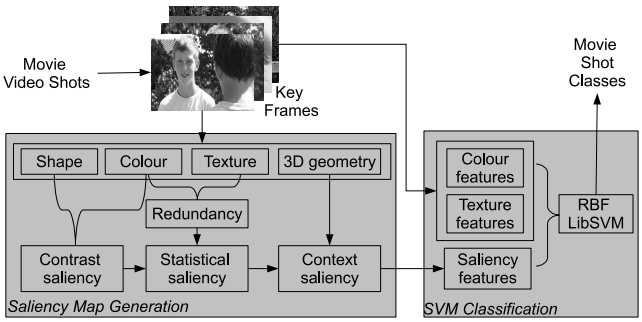

Fig. 1. The framework of context saliency based movie shot classification

3. MOVIE SHOT CLASSES

Fig. 2 shows the eight most common types of video shots in movies. *Wide-Shot* consists of Extreme Wide Shot, Very Wide Shot and Wide Shot. *Mid-Shot* shows a part of the subject in more detail while still giving an impression of the whole subject. *Close-Up* is that a certain feature or a part of the subject takes up the whole frame. *Extreme-Close-Up* gets right in and shows extreme detail. *Cut-In* shows some (other) parts of the subject in detail. *Cut-Away* is a shot of something other than the subject. *Two-Shot* is a shot of two people, framed similarly to a mid shot. *Over-the-Shoulder-Shot* is looking from behind a person at the subject. Since *Extreme-Close-Up*s are extremely rare in movies, we consider only 7 classes of shots in this paper.

4. CONTEXT SALIENCY MAP GENERATION

Generally speaking, movie shots are designed to make movie characters easily attracting audience attention. Special cases are: *Wide-Shot*s shooting scenes, *Cut-In*s and some *Close-Up*s on an object when a particular part of a character or an object is important to the movie story. In most cases, characters or significant objects, which are related to movie story, belong to foreground in movie frames. On the other hand, being foreground, such as trees, does not mean to be

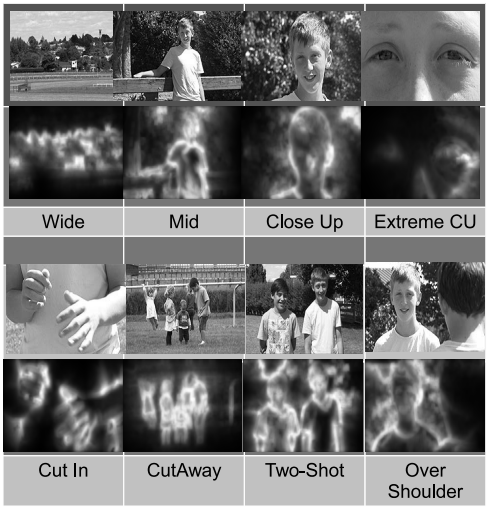

Fig. 2. The most common shot types in movies

attracting audience attention or significant to movie story. In order to generate effective saliency maps for movie shot classification, we need to consider two issues. 1) The salient region is more likely to be foreground rather than background. Unimportant background should have low saliency. 2) The salient regions should be statistically rare. Regions from frequently appeared (unimportant) objects and background should have low saliency. Unimportant objects, such as tree, grass and mountain, even detected as foreground should have low saliency. Therefore, redundancy analysis and geometric constraint are introduced.

Firstly, multi-scale contrast saliency $S(x, P)$ is calculated, based on color and shape features, using a linear combination of contrasts in the Gaussian image pyramid [16]. A face detection module is added to improve the saliency map with the tree-structured multiview face detector (MVFD) [17]. $S'(x, P) = S(x, P) + \sum_{k=1}^{N} \pi_k N(p_k, v_k)$, where p_k is the center of detected face. v_k is corresponding variances.

Secondly, redundancy analysis modifies the contrast saliency in order to get statistical saliency. 50 wide shot indoor/outdoor images, labeled with foreground and background, are selected. Each selected image is divided into $9px \times 9px$ patches. Color histogram and gray level co-occurrence matrix are extracted from each patch. K-mean clustering is applied on all the patches in 50 images to get 7 clusters. By checking the patches, we find that 7 clusters represent sky, cloud, water, sand, ground, grass and tree roughly. These patches are later used as sample patches to calculate information density as follows. The keyframes are also divided into $9px \times 9px$ patches. Let:

$$S_r = -log_2 \frac{S_{r-1}}{\min\{Dis(h_r, h_s)\}} \times S'_r(x, P) \qquad (1)$$

, where S_{r-1} and S_r denote saliency of two continuous patches. $Dis(h_r, h_s)$ is the color histogram distances be-

tween current patch and sample patches. S_r is then normalized by $S'_r = \frac{S_r}{\max(S_r)}$. In order to reduce the redundancy and blob noise, the patches with low information density are removed from contrast saliency map.

Thirdly, we adopt the geometric context from a single image [18] to extract geometric information. It estimates the coarse geometric properties of a scene by learning appearance-based models of geometric classes with a multiple hypothesis framework. Superpixel label confidences, weighted by the homogeneity likelihood, are determined by averaging the confidence in each geometric label of the corresponding regions:

$$G(y_i = v \mid x) = \sum_{j}^{n_h} P(y_j = v \mid x, h_{ji}) P(h_{ji} \mid x) \quad (2)$$

where G is the label confidence, y_i is the superpixel label, v is a possible label value, x is the image data, n_h is the number of hypotheses and h_{ji} defines the region containing the i^{th} superpixel for the j^{th} hypothesis with the region label y_j.

With statistical saliency $S(x, y)$ and geometric constraint $G(x, y)$, context saliency $C(s)$ could be modeled by Bayesian framework using the maximum a posteriori (MAP) criterion. Fig. 2 shows the context saliency maps generated from the example images in the upper row.

5. SVM CLASSIFICATION

In order to identify movie shot types, features are extracted from saliency maps for SVM classification.

5.1. Feature Extraction

We extract 80 Dimension features from both context saliency maps and the original keyframes to describe global and local characteristics.

Firstly, features representing the distribution of salient regions are extracted from context saliency maps. One saliency map is divided into 16 grids equally. The saliency value is calculated for each grid to get 16 dimensions of features.

Secondly, gray histogram is calculated with 32 bins from the original keyframes.

Finally, each keyframe is divided into 16 grids. Texture co-occurrence matrix is calculated to get the entropy and the correlation for each grid.

5.2. Classification

Support Vector Machine (SVM) is applied for classification. We notice that shot types are very unbalanced in movies. For example, some shot types, such as *Wide-Shot* and *Cut-In*, rarely appear in movies. Some types appear quite often, such as *Close-Up*. To deal with the unbalanced data, we choose the radial basis function (RBF) as a kernel function, i.e., $K(x_i, x_j) = exp(-r||x_i - x_j||^2)$, $r > 0$, for SVM classification. One-against-all multi-class approach is used.

Through a kernel, the training data are implicitly mapped from a feature space to a kernel space, and an optimal separating hyperplane is determined therein. Let R denote the n-dimensional feature space. The training data are $(x_i, y_i), i = 1, \ldots, l$ where $x_i \in R^n$, $y \in \{-1, +1\}^l$. SVM finds an optimal separating hyperplane that classifies the one class against all by the minimum expected test error. The hyperplane has the following form: $\langle w, x_i \rangle + b = 0$, where w and b are the normal vector and bias, respectively. The training feature vectors x_i are mapped into a higher dimensional space using the function ϕ. The problem of finding the optimal hyperplane is a quadratic programming problem of the following form [19]:

$$\min_{w,\xi} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i \quad (3)$$

with the constraints $y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i, \quad i = 1, 2, \ldots, l$. C is the cost controlling the trade-off between function complexity and training error, and ξ_i is the slack variable.

6. EXPERIMENTAL RESULTS

The experiments were performed on 3206 movie shots collected to cover four movies genres including Action, Horror, Comedy and Drama. Ground truth of shot classes was manually labeled. *Extreme-Close-Up*s were ignored for classification since they were extremely rare in our dataset. We applied a 3-fold cross validation. Fig. 3 shows the Precision-Recall curve for classifying seven shot types. Classification for *Wide-Shot* and *Cut-In* are not satisfactory. The main reason is that the samples of these two classes are less 2% of the total samples.

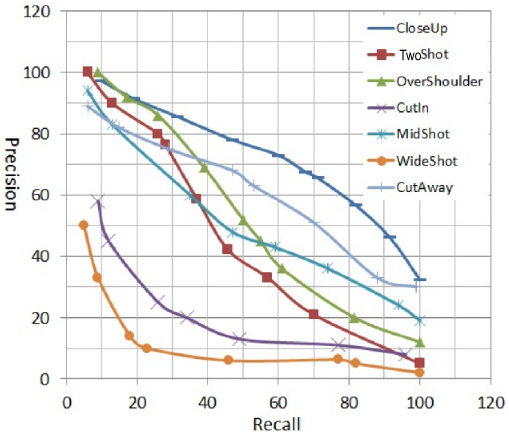

Fig. 3. Precision-Recall curve for 7 shot types

In order to justify that context saliency plays an important role in classification. We classify *Close-Up*, the most popular shot type in movies, using three different feature sets, which

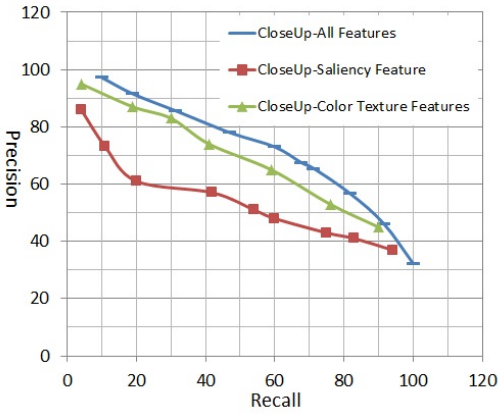

Fig. 4. Precision-Recall curve for Close Up using different feature set

are the sets of all features, saliency features and color texture features. From Fig. 4, it is obvious that saliency features improve movie shot classification significantly.

Table 1 shows the F-measure for classification of 7 shot types. From Table 1, we find that the F-measure of *Close-Up*, *Cut-Away* and *Over-the-Shoulder-Shot* are higher than the rest shot types. These three are the top three popular shots in our dataset while *Wide-Shot* and *Cut-In* are very rare (less than 2%). It should be noted that the ratio of the samples within dataset heavily affects the degree of classification difficulty. The smaller the ratio is, the more difficult the classification is. Another possible issue which affects classification is data labeling. While manually labeling the data, we find it is not easy to distinguish *Two-Shot* from *Mid-Shot* for some cases. In movies, these two shots are actually framed similarly. If the depth between two persons is big within a keyframe, the corresponding shot should be labeled as *Mid-Shot*. Mis-labeling of *Mid-Shot* and *Two-Shot* somehow affects the classification of these two shot types.

Table 1. F-measure for 7 shot types classification

Shot Type	CU	TS	OS	CI	MS	WS	CA
F-measure	68.2	43.8	51	25.5	47.5	15.8	60

Note: CU: CloseUp; TS: TwoShot; OS: OverShoulder; CI: CutIn; MS: MidShot; WS: WideShot; CA: CutAway.

7. CONCLUSIONS

Movie shot classification is a vital and challenging task. In this paper, we have proposed a context saliency based movie shot classification method. The context saliency has been produced by removing redundancies with low information densities from the contrast saliency and incorporating geometry constrains. Compared to the traditional saliency map, the context saliency accurately represents the visual attention distributed in a video frame. Through experiments, the context saliency has been proved to be significant for shot classification although unbalanced movie shots have somehow affected classification.

8. REFERENCES

[1] D.-H. Wang, Q. Tian, S. Gao, and W.-K. Sung, "News sports video shot classification with sports play field and motion features," in *Proceedings of International Conference on Image Processing*, 2004, vol. 4, pp. 2247–2250.

[2] M.-C. Tien, H.-T. Chen, Y.-W. Chen, M.-H. Hsiao, and S.-Y. Lee, "Shot classification of basketball videos and its application in shooting position extraction," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007, vol. 1, pp. I–1085 – I–1088.

[3] C. G. M. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," *Multimedia Tools and Application*, vol. 25, no. 1, pp. 5–35, 2005.

[4] M. Lazarescu and S. Venkatesh, "Using camera motion to identify types of american football plays," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2003, vol. 2, pp. 181–184.

[5] L.-Y. Duan, M. Xu, Q. Tian, C.-S. Xu, and J. S. Jin, "A unified framework for semantic shot classification in sports video," *IEEE Transactions on Multimedia*, vol. 7, no. 6, pp. 1066–1083, 2005.

[6] J. Fan, A. K. Elmagarmid, W. G. Aref X. Zhu, and L. Wu, "Classview: hierarchical video shot classification, indexing, and accessing," *IEEE Transactions on Multimedia*, vol. 6, no. 1, pp. 70–86, 2004.

[7] C. Lang, D. Xu, and Y. Jiang, "Shot type classification in sports video based on visual attention," in *Proceedings of the International Conference on Computational Intelligence and Natural Computing*, 2009, vol. 1, pp. 336–339.

[8] H. L. Wang and L.-F. Cheong, "Film shot classification using directing semantics," in *Proceedings of the International Conference on Pattern Recognition*, 2008, pp. 1–4.

[9] H. L. Wang and L.-F. Cheong, "Taxonomy of directing semantic for film shot classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 10, pp. 1529–1542, 2009.

[10] S. Wang, S. Jiang, Q. Huang, and W. Gao, "Shot classification for action movies bsed on motion characteristics," in *Proceedings of the IEEE International Conference on Image Processing*, 2008, pp. 2508–2511.

[11] "Shot types," http://www.mediacollege.com/video/shots/.

[12] L. Yang, B. Geng, A. Hanjalic, and X. S. Hua, "Contextual image retrieval model," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2010, pp. 406–413.

[13] Y.-F. Ma, X.-S. Hua, L. Lu, and H. J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Transactions on Multimedia*, vol. 7, no. 5, pp. 907–919, 2005.

[14] H. Liu, M. Xu, Q. Huang, J. S. Jin, S. Jiang, and C. S. Xu, "A close-up detection method for movies," in *Proceedings of the IEEE International Conference on Image Processing*, 2010, pp. 1505–1508.

[15] L. Shi, J. Wang, H. Lu, and C. S. Xu, "Context saliency based image summarization," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2009, pp. 270–273.

[16] J. Sun T. Liu, N. N. Zheng, X. Tang, and H. Y. Shum, "Learning to detect a salient object," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[17] C. Huang, H. Ai, Y. Li, and S. Lao, "Vector boosting for rotation invariant multi-view face detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2005, vol. 1.

[18] D. Hoiem, A. A. Efros, and M. Hebert, "Geometric context from a single image," in *ICCV*, 2005.

[19] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.