Conceptual indexing of television images based on face and caption sizes and locations.

Remi Ronfard *, Christophe Garcia †, Jean Carrive*

*INA, 4 avenue de l'Europe, 94366, Bry-sur-Marne, France
†ICS–FORTH, P.O.Box 1385, GR 711 10 Heraklion, Crete, Greece
Email: {rronfard,jcarrive}@ina.fr, cgarcia@ics.forth.gr

Abstract Indexing videos by their image content is an important issue for digital audiovisual archives. While much work has been devoted to classification and indexing methods based on perceptual qualities of images, such as color, shape and texture, there is also a need for classification and indexing of some structural properties of images. In this paper, we present some methods for image classification in video, based on the presence, size and location of faces and captions. We argue that such classifications are highly domain-dependent, and are best handled using flexible knowledge management systems (in our case, a description logics).

1 Introduction

Classifying shots based on their visual content is an important step toward higher-level segmentation of a video into meaningful units such as *stories* in broadcast news or *scenes* in comedy and drama. Earlier work on the subject has shown that shot similarity based on global features such as duration and color could be efficient in limited cases [14,1]. More recent work tends to highlight the limits of such techniques, and to emphasize more specific features, such as caption and face sizes and locations [11,12,9].

Captions and faces are powerful video indexes, given that they give generally a clue about the video content. In video segmentation, they may help to find program boundaries, by detecting script lines and to select more meaningful keyframes containing textual data and/or himan faces. Automatic detection of programs, such as TV Commercials or news, becomes possible using location and size of text.

One important issue that is not dealt with by previous work is the necessity of exploiting domain knowledge, which may only be available at run-time. In this paper, we establish a clear-cut separation between *feature extraction* which is based on generic tools (face detection, caption detection) and *classification*, which is based on heuristic, domain-specific rules. With examples drawn from real broadcast news , we ilustrate how such classes can be organized into taxonomies, and used as indexes in large audiovisual collections.

2 Description logic databases.

We use the CLASSIC Description Logics system [2] as a representation for both the image *classes* and the image *observations*, which are obtained through video analysis. CLASSIC represents classes as *concepts* which can be *primitive* or *defined*. Primitive concepts are only represented with necessary conditions. We use them to represent event classes which are directly observable : shots, keyframes, faces and captions. The necessary conditions determine the inferences which can be drawn in such classes : for instance, shot have at least one keyframe, keyframes may have faces or captions. Defined concepts are represented with both necessary and sufficient conditions. Therefore, class membership can be inferred automatically for defined concepts. In this paper, we focus on defined concepts for keyframe and shot classes. Relations between concepts are called *roles*, and one important role between audiovisual events is containment (*part-of* role). Concepts and roles are organized in taxonomies, such as the one shown in Fig.1, which contains both primitive and defined concepts implemented in our current prototype.

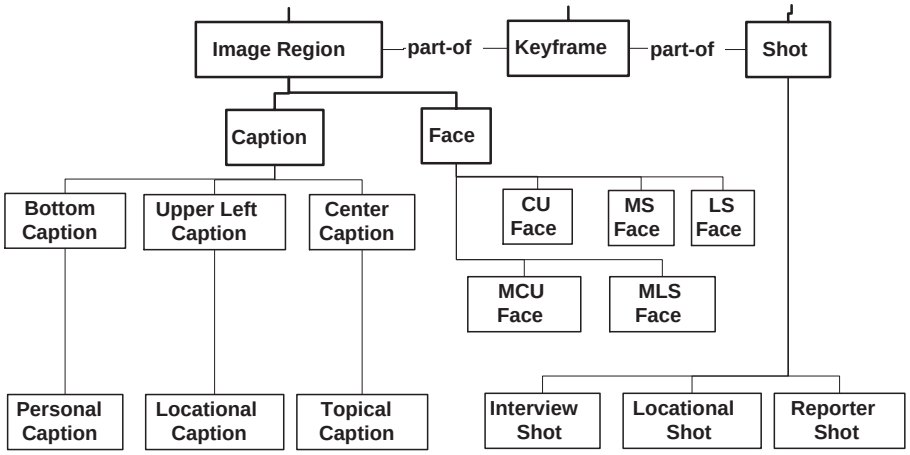

Figure1. A taxonomy of image regions, keyframes and shots. Context-specific classes are defined in terms of more generic classes using subsumption and part-of links.

3 Feature extraction

In general, a shot can be represented synthetically by a small number of static keyframes. We select keyframes by clustering them based on their color content, and deciding on the right number of clusters, based on a local test performed

during shot detection. The video segmentation and keyframe extraction tools of DIVAN have been described elsewhere [3], and we focus here on the techniques used to detect faces and captions.

3.1 Face detection

Faces appearing in video frames are detected using a novel and efficient method that we presented in details in [7]. The proposed scheme is designed for human faces detection in color images under non-constrained scene conditions, such as the presence of a complex background and uncontrolled illumination. Color clustering and filtering using approximations of the HSV skin color subspaces are applied on the original image, providing quantized skin color regions which are iteratively merged in order to provide a set of candidate face areas. Constraints related to shape and face texture analysis are applied, by performing a wavelet packet decomposition on each face area candidate and extracting simple statistical features such as standard deviation. Compact and meaningful feature vectors are built with these statistical features. Then, the Bhattacharrya distance is used for classifying the feature vectors into face or non-face areas, using some prototype face area vectors, acquired in a previous training stage. For a data set of 100 images with 104 faces covering most of the cases of human faces appearance, a 94.23% good detection rate, 20 false alarms and a 5.76% false dismissals rate were obtained.

3.2 Caption detection

Our method for caption detection is especially designed for being applied to the difficult case where text is superimposed on color images with complicated background and is described in [8]. Our goal is to minimize the number of false alarms and to binarize efficiently the detected text areas so that they can be processed by standard OCR software. First, potential areas of text are detected by enhancement and clustering processes, considering most of constraints related to the texture of words. Then, classification and binarization of potential text areas are achieved in a single scheme performing color quantization and characters periodicity analysis. First results using a data set of 200 images containing 480 lines of text with character sizes ranging from 8 to 30, are very encouraging. Our algorithm detected 93% of the lines and binarize them with an estimated good readability rate of 82%. An overall number of 23 false alarms have been found, in areas with contrasted repetitive texture.

4 Shot classification

The automatic detection of human faces and textual information provides users with powerful indexing capacities of the video material. Frames containing detected faces or text areas may be searched according to the number, the sizes, or the positions of these features, looking for specific classes of scenes. Number

Figure2. Some results of face and text detection.

and size of detected faces may characterize a big audience (multiple faces), an interview (two medium size faces) or a close-up view of a speaker (a large size face). Location and size of text areas helps in characterizing the video content especially in news. In this section, we explain in more details how such shot classes can be defined, and their instances recognized automatically, in a DL framework.

4.1 Face classes

The first axis for shot classification is the apparent size of the detected faces. Faces are an important semantic marker in images, and they also serve as a very intuitive and immediate spatial reference. With respect to the human figure, cinematographers use a vocabulary of common framings, called 'shot values' from which we have selected five classes, corrresponding to cases where the face can be seen and detected clearly. They range from the close-up (CU), where the face occupies approximately half of the screen, to the long shot (LS), where the human figure is seen entirely, and the face occupies around ten percent of the

screen. Intermediate shot values are the medium shot (MS), the medium-close-up (MCU) and the medium-long-shot (MLS) [13].

Shot value classes are usually defined in relative and imprecise terms, based on the distance of the subject to the camera. In order to provide a quantitative definition, we use the fact that in television and film, the apparent size of faces on the screen vary inversely with their distance to the camera (perspective shortening). We therefore compute the quantity $d = \frac{FrameWidth}{FaceWidth}$ and classifiy the face regions according to five overlapping bins, based on a uniform quantization of d in the range of $[0, 12]$ (see Table 1). Note that this is consistent with the resolution used (MPEG-1 video with 22 macroblocks per line).

The five face classes shown in Fig.1 follow immediately from the correspondance shown in Table 1. Given such classes, it is possible to define keyframe classes based on the number and size of their detected faces. When all faces are in a given class (mcu-face) then the keyframe itself can be qualified (mcu-frame). Note that in the case of multiple face classes, we do not attempt to classify the keyframe. But using overlapping face value classes allows us to automatically classify the frame into the common class of all its detected faces, in most practical cases.

Value	CU	MCU	MS	MLS	LS
Size	1/2	1/4	1/6	1/8	1/10
Range	$d \leq 4$	$2 \leq d \leq 6$	$4 \leq d \leq 8$	$6 \leq d \leq 10$	$8 \leq d$

Table1. Face sizes, distances and shot values.

4.2 Caption classes

While faces are classified according to their dimension, captions are best classified according to their position on the screen. In many contexts, such as broadcast news, the caption location determines the semantic class of the caption text. As an example, Fig.2 shows examples of three caption classes : topical (center-left caption), personal (bottom caption) and locational (upper-left caption). In this case, we therefore define three caption classes based on simple geometric tests for bottom, upper-left and center-left captions, as we did with faces. But we propagate the class memberships from captions to frames and shots in a very different way from what did with shot values, because in this case the presence of a single center-left caption suffices to classify the frame as a *topical keyframe*, and the shot as a *topical shot*. Since CLASSIC does not provide the existential operator, this is done with a special-purpose propagation rule, triggered for all center-left captions.

4.3 Shot classes

Shot classification immediately follows from keyframe classification in the case of *simple shots* (shots with exactly one keyframe). Shots containing more than one keyframe are qualified as *composite shots* and are only classified as CU, MCU, etc. when *all* their keyframes are in the same class. In all other cases, we leave them unclassified, for lack of more specific information. Curiously, this limitation coincides with limitations of CLASSIC itself, which can only handle conjonctions of role restrictions, but not negations or disjunctions. In the future, we will investigate other DL systems to overcome this limitation. As another extension, we are currently developping a constraint-based temporal reasoning system on top of CLASSIC, which will allow us to define and classify a composite shot as a *zoom in from MS to CU* [4].

In some contexts, such as broadcast news or sports, more specialized shot classes can be defined, using simple combinations of the previously introduced classes. For instance, an *interview shot* can be defined as a one-shot which is both an MCU-shot and a personal-shot. A reporter shot can be defined similarly, as a one-shot, MCU, locational shot. And an anchor shot can be defined as a one-shot, MCU, topical shot. While such classes are only valid within a particular context, they allow useful inferences, especially when dealing with large collections of very similar television broadcasts.

5 Experimental results and further work

Our shot classification system has been tested as part of the DiVAN prototype. DiVAN is a ditributed audiovisual archive network which uses advanced video segmentation techniques to facilitate the task of documentalists, who annotate the video contents with time-coded descriptions. In our experiments, the video is processed sequentially, from segmentation to feature extraction, to shot classification and scene groupings, without human intervention, based on a precompiled shot taxonomy representing the available knowledge about a collection of related television programs.

S1	S2	S3	S4	S5	S6
CU	MLS, MS	MCU, MS, Topical, Anchor	CU, MCU, Personal, Interview	MCU, MS, Locational, Presonal, Reporter, Interview	CU, MCU, Personal, Interview

Table2. Shot classification results for Fig.2

In Fig.2, we present some results of the proposed face and text detection algorithms. The first line shows typical examples of multiple faces in the scene

and close-up view of a face. The other examples illustrate the case of face and text detection appearing in the same frame. Table 2 shows the classification results for those shots. In those examples, it should be noted that multiple or even conflicting interpretations (such as Interview and Reporter shot) are allowed. We believe that such ambiguities can only be resolved by adding more knowledge and more features into the system.

One way of adding such knowledge is to go from detection to recognition. Face and caption recognition enable more powerful indexing capacities, such as indexing sports programs by score figures and player names, or indexing news by person and place names. When detected faces are recognized and associated automatically with textual information like in the systems Name-it [10] or Piction [5], potential applications such as news video viewer providing description of the displayed faces, news text browser giving facial information, or automated video annotation generators for faces are possible.

In order to implement such capabilities, we are developing an algorithm dedicated to face recognition when faces are large enough and in a semi-frontal position [6]. This algorithm uses directly the features extracted in the detection stage. As an addition, our algorithm for text detection [8] includes a text binarization stage that makes the use of standard OCR software possible. We are also currently completing our study by using a standard OCR software for text recognition. With those capabilities, we will be able to extend the number of shot classes recognized by our system, to recognize shot sequences, such as shot-reverse-shots, and to resolve ambiguous cases, such as determining whether two keyframes contain the same faces or not (within a shot boundary).

6 Conclusions

Based on extracted faces and captions, we have been able to build some useful classes for describing television images. The description logic framework used allows us to easily specialize and extend the taxonomies. Classification of new instances is performed using a combination of numerical methods and symbolic reasoning, and allows us to always store the *most specific* descriptions for shots or groups of shots, based on the available knowledge and feature-based information.

References

1. Aigrain, Ph., Joly, Ph. and Longueville, V. Medium knowledge-based macro-segmentation of video into sequences Intelligent multimedia information retrieval, AAAI Press - MIT Press, 1997.
2. Borgida, A., Brachman, R.J., McGuiness, D.L., Resnick, L.A. 1989. CLASSIC: A Structural Data Model for Objects. ACM SIGMOD Int. Conf. on Management of Data, 1989.
3. Bouthemy, P., Garcia C. , Ronfard R. , Tziritas G. , Veneau E. Scene segmentation and image feature extraction for video indexing and retrieval. VISUAL'99, Amsterdam, 1999.

4. Carrive, J., Pachet F. , Ronfard R. Using Description Logics for Indexing Audiovisual Documents. Proceedings of the International Workshop on Description Logics, Trento, Italy, 1998.
5. Chopra K. , Srihari R.K. . Control Structures for Incorporating Picture-Specific Context in Image Interpretation. in: *Proceedings of Int'l Joint Conf. on Artificial Intelligence*, 1995.
6. Garcia C. , Zikos G. , Tziritas G. . Wavelet Packet Analysis for Face Recognition. To appear in *Image and Vision Computing*, 18(4).
7. Garcia C. and Tziritas G. . Face Detection Using Quantized Skin Color Regions Merging and Wavelet Packet Analysis. *IEEE Transactions on Multimedia*, 1(3):264–277, Sept. 1999.
8. Garcia C. , Apostolidis X. . Text Detection and Segmentation in Complex Color Images. *IEEE International Conference on Acoustics, Speech, and Signal*, June 5-9 2000, Istanbul, Turkey.
9. Ide, I., Yamamoto, K. and Tanaka, H. Automatic indexing to video based on shot classification. Advanced Multimedia Content Processing, LNCS 1554, November 1998.
10. Satoh S. , Kanade T. . Name-it: Association of Face and Name in Video. in: *Proc. of Computer Vision and Pattern Recognition. IEEE Compu ter Society Press*, pp. 368-373, 1997.
11. Gunsel, B. and Ferman, A.M. and Tekalp, A.M. Video Indexing Through Integration of Syntactic and Semantic Features. WACV, 1996.
12. Ferman, A.M., Tekalp, A.M. and Mehrotra, R. Effective Content Representation for Video IEEE Intern. Conference on Image Processing, October 1998.
13. Thomson, R. Grammar of the shot. Media Manual, Focal Press, Oxford, UK, 1998.
14. Yeung, M. and Yeo, B.-L. Time-constrained Clustering for Segmentation of Video into Story Units International Conference on Pattern Recognition, 1996.