Improvement of a Person Labelling Method Using Extracted Knowledge on Costume

Gaël Jaffré and Philippe Joly

Université Paul Sabatier,
IRIT - Équipe SAMOVA,
31062 Toulouse Cedex 09,
France
{jaffre, joly}@irit.fr

Abstract. This paper presents a novel approach for automatic person labelling in video sequences using costumes. The person recognition is carried out by extracting the costumes of all the persons who appear in the video. Then, their reappearance in subsequent frames is performed by searching the reappearance of their costume. Our contribution in this paper is a new approach for costume detection, without face detection, that allows the localization of costumes even if persons are not facing the camera. Actually face detection is also used because it presents a very accurate heuristic for costume detection, but in addition in each shot mean shift costume localization is carried out with the most relevant costume when face detection fails. Results are presented with TV broadcasts.

1 Introduction

Our framework is the analysis of costume as a feature for video content indexing, and especially its automatic extraction. Some experiments made on automatic video summarization showed that the costume feature is one of the most significant clue for the identification of keyframes belonging to some given excerpt [1]. Authors justify this property by the fact that costumes are attached to character function in the video document. Costume is already used as an entity for audiovisual production description scheme [2,3], but only for a theoretical point of view, without automatic detection. Only recently an automatic application using costume was introduced [4].

However, the costume detection remains a problem, because at the moment it is only based on face detection, and so is dependant of the face detector and fails when the faces are too small in the frame. We can find papers in literature where clothes are used to help the recognition [5,6], but in each of them the costume detection is based on face detection. Our contribution in this paper is a new approach for costume detection, without face detection, that allows the localization of costumes even if persons are not facing the camera. Actually face detection is also used because it presents a very accurate heuristic for costume

A. Gagalowicz and W. Philips (Eds.): CAIP 2005, LNCS 3691, pp. 489–497, 2005.

Class 1 Class 2 Class 3

Fig. 1. Classification of character framings

detection, but in addition in each shot, when face detection fails, mean shift costume localization is carried out with the most relevant costume.

In section 2 we introduce the application of person labelling using costumes. Section 3 presents the costume detection algorithm. Results are presented in section 4.

2 Person Labelling Using Costume

The goal of this application is to automatically create an index which gives, for each frame, all the persons who are present. The application described in this paper is automatic: the first time a character appears, it is added in a costume database with an automatic label. At the end of the processing, the user can update the index by giving a real name to each label.

2.1 Concepts of Shot and Character Framing

In this application, we use the notion of shot. It roughly corresponds to a set of continuous frames taken with an uninterrupted recording of a video camera. As we work on video sequences extracted from TV talk shows, there are only slight camera motions during a same shot, and no person appear or disappear during a shot, the number of persons remains constant. So, in a same shot we can run the costume detection only with some frames, and generalize the results with the remaining frames. Fig. 2 presents examples of propagations.

We call "character framing" the significance of the person according to his position and size in the frame. We considered three classes of framing: the first one corresponds to a character who is centered, and has a sufficient size to be the most important visual interest in the frame. The second one corresponds to characters who are important components of the frame, among several others. The third one corresponds to background characters, or characters who are not easily identifiable. Fig. 1 shows an example. This classification will be significant for the shot propagation (section 2.2), and for the experiment part (section 4).

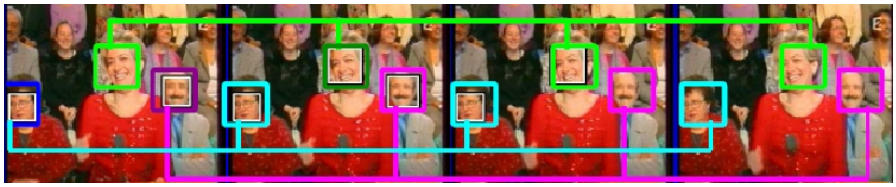

Fig. 2. Propagation of person detections. The white and black boxes represent the automatic detection provided by the face detector. For each color, the dark box represents the validated faces, the others are the propagated faces. Actually the faces are not directly propagated, this example is only here to better understand the principle, in our real application we only propagate the labels of the detected characters.

2.2 Algorithm of the Application

The goal of the application is to detect and recognize all the persons that appear in each frame. The following algorithm is applied in each shot of the video.

The first step is the detection of faces in the first frames of the shot. Character framing have here an interest: if a first class face is detected (at the moment we consider a centered face as a first class face), then we consider that we have detected the only useful information, so we stop the detection, and propagate (backward and forward) the results to all the frames of the shot. If no face is detected, or only second or third class faces, the search goes on within next frames, because we consider that we could have missed some faces. If this new search does not provide any face, then we consider that the face detector failed.

If faces are detected at any step, then the costume of each person is extracted (from the frame where the face was detected) according to the face locations. The features of each costume are extracted, and compared to the ones of the database. If a costume corresponds, then the person wearing it is recognized. Else, the new costume is added in the database with an automatic label. In both cases this person is considered as present in all the frames of the shot.

When the face detector fails, we add a new step, which is costume localization without face detection. This step will be detailed in section 3.3. With this additional step we can deal with the frames where the face is not detected. Due to computational time, this detection will be carried out in only one frame. If no costume is detected in spite of this step, we finally consider that no character is present in the shot.

2.3 Shot Boundary Detection

Shot boundary detection can be a challenging task, if the boundaries are gradual. However as our application process only TV talk-shows, we do not have the problem of gradual transitions, because the transitions have at most two frames, so a very simple detector is sufficient. Moreover, we need a very fast preprocessing tool, providing exploitable results with a minimum cost of the system resources,

in order to keep some for the costume processing, and be able to have real-time processing on a modest computer.

We subsample the frame with a ratio of eight (for both rows and columns), and we take only one channel out of three. Then, we compare each pixel to the same pixel in the previous frame. We consider that the two frames belong to different shots if the mean difference is over a threshold. Under this threshold we consider that the two frames are in the same shot. This algorithm allows exploitable results on our kind of contents, with a very fast processing.

3 Costume Detection

We can find many methods in literature to detect people presence in images, however there are all focused on some special content. First, pedestrian detection focuses on detecting persons, but the context of the applications is often for future driving assistance systems [7], with specific conditions. Some applications dedicated to surveillance allow the detection of persons with different scales [8], but under restricting hypothesis, like fixed video cameras. These methods would not be usable for our application, because our video corpus contains various framings, such as close shots, as well as global views, with mobile cameras. Moreover, it is very common that the whole body does not appear in the frame, just the upper part, which is problematic for these methods.

3.1 Face-Based Costume Detection

Recently, face was used as a visual clue for person detection [4,6]. The main idea is the use of face detection algorithms to detect human presence. Nowadays, face detection is not yet a solved problem, but the existing algorithms produce good results when the input images are not very complex, which is often the case in our corpus of TV broadcasts.

Thus, the first step of our costume detection is the run of a face detection algorithm, so as to detect the different possible characters who are present in the current frame, and their approximate position and scale. Then, the costume of each character is extracted from the image according to the location and the scale of his face.

There are many methods for face detection in literature (see [9] for a recent review), but we do not use a specific one. We intend to make an application which is independent of the face detector, when this one is able to produce some results of at least a given minimal quality. We used the method presented in [10], because a fast implementation is available in the Intel library OpenCV [11].

The costumes are extracted according to the localization and the scale of the detected faces. At the moment, we estimate the costume by the area under the face. The size of this area is proportional to the one of the face. In our examples, we used a width size of 2.3 times the one of the face, and for the height size a ratio of 2.6. We chose experimentally these coefficients by taking the ones which give the best fitting of the box in our learning images.

3.2 Face Detection Improvement

The algorithm of costume localization is based upon face detection. However, frame by frame face localization introduce many false alarms, due to some noise present in the data. Only one false detection in a frame is enough to involve a false alarm on costume detection.

In order to reduce these false detections, we must exploit the properties of a video sequence by using a temporal approach. For each frame, we detect all the faces using a static approach. Then, we take a temporal window (subsequence) of $2N+1$ frames. For each candidate face, we count its number of occurrences in the N previous frames, and in the N next frames. Recall that all these detections are made independently. Then, we keep a candidate face if it appears at least N_2 times in this subsequence. In our application, we took $N = 2$ (which leads to a subsequence of 5 frames) and $N_2 = 4$.

We consider that two detected faces correspond to the same face if there are roughly at the same location. The position parameters may slightly vary considering camera works or character motions. So, a small variation of these parameters is borne to take into account these effects. Moreover, to avoid the detection of faces in dissolves we consider that two faces correspond to the same face if the costumes detected from these faces are also identical (in terms of features, cf section 3.4).

3.3 When the Face Detector Fails

Even if face detection is robustified (cf. section 3.2), there are many frames where the face is occluded, where the person is shot from behind, or where the face detection fails. In order to deal with the case where the persons are not detected using face detection, we added a costume detection step which is not based on face detection.

Costume Classification. Unlike face-based costume detection, we do not have any prior information about the costume location in the frame. So, searching for each model of costume can be very computationally expensive. In order to reduce this cost, we will only search for the costume which is the most likely to be in the frame.

We suppose that if a costume is present in a frame with the same scale, then its histogram h_c is included in the histogram of the frame h_f. So, the histogram intersection [12] with non-normalized histograms would provide as a result the costume histogram h_c

$$\sum_{i=1}^{n} \min\left(h_c^i, h_f^i\right) = \sum_{i=1}^{n} h_c^i \tag{1}$$

So as to deal with the case where the costume does not have the same scale in the frame, and to obtain a fractional match value between 0 and 1, the intersection is normalized by the number of pixels in the model histogram, and compared to the sum of the costume histogram. So for each costume the coefficient $C_{h_f}(h_c)$ is computed by

$$C_{h_f}(h_c) = \frac{\sum_{i=1}^{n} h_c^i - \sum_{i=1}^{n} \min\left(h_c^i, h_f^i\right)}{\sum_{i=1}^{n} h_c^i} \tag{2}$$

Each costume is tested to see if its colors are present in the frame, and then the costumes of the database are sorted by relevant color. Then, we only search in the frame the localization of the most relevant costume.

Costume Localization. Now we have a unique model of costume to find in the frame, the problem reduces to detect its presence or not in the frame, and if so to find its location. To quickly find its location using only its color histogram, we use the object detection approach presented in [13]: using the costume histogram, an image of weights is created from the frame, which represents the repartition of the most probable pixels to be part of the object. This image of weights is called backprojected image, and is based on the ratio histogram [12] $r_k = \min\left(\frac{h_c}{h_f}, 1\right)$.

Since the ratio histogram emphasizes the predominant colors of the costume while diminishing the presence of clutter and background colors, the backprojected image represents a spatial measure of the costume presence.

From this image of weights, the problem is to find if there is a "group" of likely pixels, and if so to detect it. Considering this image as a cluster in $\mathbb{R}^2$, the "group" of pixels can be considered as the cluster global mode. Then, a statistical method, the mean shift procedure [14], is used to detect it.

If we note $\{\mathbf{x}_i\}_{i=1\ldots n}$ the set of points of the cluster, and $w(\mathbf{x}_i)$ the weight associated to pixel $\mathbf{x}_i$, then the mean shift vector for the point $\mathbf{x}$ is computed by

$$M_h(\mathbf{x}) = \frac{\sum_{\mathbf{x}_i \in S_h(\mathbf{x})} w(\mathbf{x}_i)\ \mathbf{x}_i}{\sum_{\mathbf{x}_i \in S_h(\mathbf{x})} w(\mathbf{x}_i)} - \mathbf{x} \tag{3}$$

where $S_h(\mathbf{x})$ is the sphere centered on $\mathbf{x}$, of radius h and containing $n_{\mathbf{x}}$ data points. More information about the mean shift procedure and mean shift vector can be found in [14]. The mean shift vector has the direction of the gradient of the density estimate at $\mathbf{x}$. The mean shift procedure is obtained by successive computations of the mean shift vector $M_h(\mathbf{x})$, and translation of the sphere $S_h(\mathbf{x})$ by $M_h(\mathbf{x})$. The procedure is guaranteed to converge [14] to a local mode. Actually, as costumes do not have the same size for height and width, we use a scale $h = (h_x, h_y)$, with $h_x > h_y$, as presented in Fig. 3.

Mean shift iterations guarantee convergence to a local mode, but we are only interested in the global mode. In order to find the global mode, we take many initializations in the frame (cf Fig. 3), and then we only keep the convergence point which brings the largest density. The density is estimated using the Parzen window [15, ch. 4]

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) \tag{4}$$

with an Epanechnikov kernel [14]

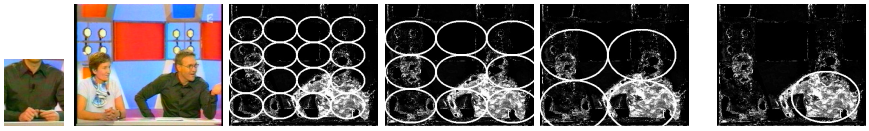

Fig. 3. Mean shift costume detection. The two first frames are the input data. The two next represent the initialization of the mean shift procedure, with the corresponding scale (h_x, h_y). The last frame is the detected costume, with the optimal scale.

$$K_E(\mathbf{x}) = \begin{cases} \frac{2}{\pi}(1 - \|\mathbf{x}\|^2) & \text{if } \|\mathbf{x}\| < 1 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

The Epanechnikov kernel was chosen because it was used to derivate the mean shift vector in equation 3 (justifications can be found in [13]).

To give up this prior information about the scale h of the costume, we run the detector many times with various scales, as shown in Fig. 3. Then, we keep the scale that provides the largest density.

Use of this Blind Approach. Using mean shift detection in addition to face-based detection can be computationally expensive if these two approaches are used in each frame, because it is carried out with various scales and several initializations. As we need mean shift detection only when the face detector fails, we apply it only one time in each shot, when the face detector provides no face in the whole shot. Thus, the processing time for blind detection is insignificant relatively to the processing time of a whole shot.

Table 1. Recognition rates for both videos

Video	Class	Number of characters	Face-based approach	+ Blind approach
1	1	19 692	18 587 (94.39%)	18 659 (94.75 %)
	2	34 978	2226 (6.34%)	2 865 (8.19 %)
	3	56 857	3 755 (6.60 %)	3 755 (6.60 %)
2	1	5 588	4 897 (87.63%)	5 005 (89.57 %)
	2	14 797	6 529 (44.12%)	6 529 (44.12 %)
	3	21 539	1 129 (5.24 %)	1 129 (5.24 %)

3.4 Similarity Measure

The feature that we use is a three-dimensional RGB color histogram. The similarity measure used to compare histograms is the Bhattacharyya coefficient, which is closely related to the Bayes error [16, p. 38]. If we note $\hat{q} = \{\hat{q}_u\}_{u=1\ldots m}$ and $\hat{p} = \{\hat{p}_u\}_{u=1\ldots m}$ the color histograms of the two costumes (m is the number of bins) the Bhattacharyya coefficient can be estimated by [17] $\rho\,(\hat{p}, \hat{q}) = \sum_{u=1}^{m} \sqrt{\hat{p}_u \hat{q}_u}$. The coefficient interval is the real interval $[0, 1]$. A value of 1 means a perfect match, whereas a value of 0 means a mismatch.

Table 2. Recognition errors for the first video. For the number of miss-classified characters, the percentage is relative to the total number of detected persons.

		Face-based approach	+ Blind approach
Video 1	false alarms	329	329
	misclassified characters	0.56%	0.54%
Video 2	false alarms	514	593
	misclassified characters	1.44%	1.43%

4 Experiments

Experiments have been carried out on different video sequences extracted from TV programs, especially TV talk-shows. We present here numerical results for two different TV talk-shows. The format of the videos is MPEG1, with a frame size of 352×288. The first video has a duration of thirty minutes, and contains 46 680 frames. The second one lasts twelve minutes, and has 18 243 frames. We manually indexed these video sequences: for each frame, we noted all the persons that appear as well as their character framing.

We compared the results for the traditional approach, only based on face detection, with our blind approach. Computational time are roughly the same for both methods: the frames were processed at a mean rate of 37 fps for the first video and 30 fps for the second one. Results are summed up in tables 1 and 2.

5 Conclusion

We proposed in this paper an approach for automatic person labelling in video sequences using costumes. We showed that on our kind of content the clothes of a person are relevant for recognition. This approach for costume detection, which is not based on face detection, allows a fast localization of the costumes when the face detector fails. We showed that results are improved when this blind approach is used in addition to face-based costume detection. However, the face-based detector is still essential, because the blind approach can only find costumes of the database, it cannot find new ones.

Moreover, we would like to significantly improve the results for the second and third class characters. A separation of the clothes in different parts (tie, jacket, hat, trousers,...) would perform a better description of the costumes, and could be used to improve the detection.

References

1. Yahiaoui, I.: Construction automatique de résumés vidéos. Thèse de doctorat, Télécom Paris, France (2003)
2. Nack, F.: AUTEUR: The Application of Video Semantics and Theme Representation for Automated Film Editing. PhD thesis, Lancaster University, UK(1996)

3. Bui Thi, M.P., Joly, P.: Describing video contents: the semiotic approach. In: Proceedings of the 2nd Content-Based Multimedia Indexing Workshop, Brescia, Italy(2001) 259–266
4. Jaffré, G., Joly, P.: Costume: A New Feature for Automatic Video Content Indexing. In: Proceedings of RIAO - Coupling approaches, coupling media and coupling languages for information retrieval, Avignon, France (2004) 314–325
5. Lerdsudwichai, C., Abdel-Mottaleb, M.: Algorithm for Multiple Faces Tracking. In: IEEE International Conference on Multimedia & Expo, Baltimore, Maryland, USA(2003)
6. Zhai, Y., Chao, X., Zhang, Y., Javed, O., Yilmaz, A., Rafi, F., Ali, S., alatas, O., Khan, S., Shah, M.: University of Central Florida at TRECVID 2004. In: Proceedings of the TRECVID 2004 Workshop, Gaithersburg, Maryland, USA(2004) 217–224
7. Broggi, A., Bertozzi, M., Chapuis, R., Chausse, F., Fascioli, A., Tibaldi, A.: Pedestrian Localization and Tracking System with Kalman Filtering. In: Proceedings of the IEEE Intelligent Vehicles Symposium, Parma, Italy(2004) 584–589
8. Yang, H.D., Lee, S.W.: Multiple Pedestrian Detection and Tracking based on Weighted Temporal Texture Features. In: Proceedings of the 17th International Conference on Pattern Recognition. Volume 4., Cambridge, UK(2004) 248–251
9. Yang, M.H., Kriegman, D., Ahuja, N.: Detecting faces in images: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **24** (2002) 34–58
10. Lienhart, R., Maydt, J.: An Extended Set of Haar-like Features for Rapid Object Detection. In: Proceedings of the IEEE International Conference on Image Processing. Volume 1., Rochester, New York, USA(2002) 900–903
11. (OpenCV) http://www.intel.com/research/mrl/research/opencv/
12. Swain, M., Ballard, D.: Color Indexing. International Journal of Computer Vision **7** (1991) 11–32
13. Jaffré, G., Crouzil, A.: Non-Rigid Object Localization from Color Model using Mean Shift. In: Proceedings of the IEEE International Conference on Image Processing. Volume 3., Barcelona, Spain(2003) 317–320
14. Comaniciu, D., Meer, P.: Mean Shift: A Robust Approach Toward Feature Space Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **24** (2002) 603–619
15. Duda, R., Hart, P., Stork, D.: Pattern Classification. Second edn. Wiley-Interscience (2001)
16. Andrews, H.: Introduction to Mathematical Techniques in Pattern Recognition. Wiley-Interscience (1972)
17. Aherne, F., Thacker, N., Rockett, P.: The Bhattacharyya Metric as an Absolute Similarity Measure for Frequency Coded Data. Kybernetika **32** (1997) 1–7