LEARNING TO IDENTIFY VIDEO SHOTS WITH PEOPLE BASED ON FACE DETECTION

Rong Jin

School of Computer Science, Carnegie Mellon University

Alexander G. Hauptmann

School of Computer Science, Carnegie Mellon University

ABSTRACT

We examine how to identify video shots with at least two humans using only detected face information. While face detection is much more reliable than shape based people classification in broadcast video, one particular difficulty is that, when there are several humans in an image, the accuracy of face detection is usually significantly degraded, which leads to poor performance in identifying shots of 'people'. Furthermore, while our standard face detector works from individual still images, we propose using the statistics of face information of images within a whole shot as additional evidence in deciding whether or not a video shot belongs to the 'people' category. Empirically, we studied which statistics of face information are more informative than others and how to combine different statistics together in order to achieve better prediction.

1. INTRODUCTION

Video understanding is an interesting and difficult problem. A standard approach is to look at low-level features to determine the content or content class of a video sequence. However, it is extremely difficult to infer a complicated concept directly from low-level image features such as color, edges, lines and textures. Therefore, it becomes desirable to first identify some middle-level general concepts and then infer more complex concepts from the identified middle-level concepts instead of trying to infer them directly from lowlevel features. In this paper, we examine one type of middle-level concept, namely the concept of 'people', which was evaluated as one of the feature extraction tasks in the video retrieval track of the TREC 2002 competition [5]. According to the definition in the TREC video track, an image belongs to the category 'people' if and only if it contains at least two humans and each of which is at least partially recognized as human. Presumably, this middlelevel concept can be very useful in building up other complicated concepts such as 'crowds', 'meetings', or 'tourists'.

Internal experiments found that while identifying people based on shape and motion is promising for surveillance footage from known cameras, this approach is not robust enough to be usable for broadcast video of the wide variety and quality in the video TREC collections of mostly documentary footage. One simplistic, but practical idea for deciding whether a video shot belongs to the category of 'people' would be to first apply a robust face detector to every image within that shot and then count the number of faces within each image. Only when there exists an image within the shot containing at least two human faces, will the whole shot be labeled as 'people'. The advantage of this simple idea is that it only requires a face detector and we don't have build any further image classification models.

One problem with this simple approach is that generally, the prediction of a face detector may not be always reliable, especially in the case of multiple faces. More specifically, when a face detector identifies multiple faces within an image, it frequently makes more mistakes than when it only identifies a single face. What happens is that the detection of multiple faces can be the (mistaken) result of some noisy textures such as stones, leafy trees or bark. Therefore, an additional classifier is required to prune out the cases of multiple faces that are misidentified by the face detector. Of course, the approach of using face information will fail completely when the people in the shot don't have identifiable faces. However, in this paper, we only focus on the idea of using face information to identify video shots of people and incorporating body shape and motion without relying exclusively on faces will be our future work.

As mentioned earlier, to go beyond identifying video shots of people by simply counting how many faces are found in each image, we need to examine other aspects of the face information provided by the face detector. Furthermore, a classifier is required to combine different aspects of face information and better predict if a shot belongs to the 'people' category. Clearly, there are two issues involved in this procedure:

- 1) Which aspects of face information will be useful in determining shots of people?
- 2) What combination model will be most effective in gluing all the evidence together in order to achieve better classification?

These two issues will be examined in Section 2.1 and Section 2.2, respectively. The empirical study will be presented in Section 3. Section 4 describes the conclusions and future work.

2. A CLASSIFICATION MODEL FOR IDENTIFYING VIDEO SHOTS WITH PEOPLE

In this section, we will first describe the set of features that can be reliably extracted from the identified face information as the basic pieces of evidence. Then, three combination models will be described that can tie all the statistics together.

2.1. Feature Extraction



Figure 1: The percentage of 'people' vs 'non-people' shots as a function of the number of images with multiple detected faces within the shot.

First, in order to illustrate how unreliable it is to identify a shot with multiple humans by simply counting the number of images with multiple faces, in Figure 1, we plot the distribution of video shots with respect to the number of images having multiple faces within a single shot for both the 'people' and 'non-people' categories. As illustrated in Figure 1, the distributions of video shots for both the 'people' and 'non-people' category are quite similar. Surprisingly, even for the 'people' category, the majority of video shots only contain images with a single identified face, namely 65% of shots in the 'people' category with only a single face. Meanwhile, in terms of the distribution of shots with images having multiple faces, both the 'people' category and the 'non-people' category display a very similar pattern though the 'people' category appears to have slightly more shots with images having multiple face than the 'non-people' category. Thus, identifying shots with people by only counting the number of faces within an image will almost be identical to a random guess. Therefore, it is important to resort to other features of face information so that further distinctions can be made between the 'people' and 'non-people' categories.

There could be two reasons why a shot with identified faces that does not belong to the 'people' category: 1) The shot contains images with only a single face and therefore does not count as a shot with 'people'. 2) The shot appears to contain images with multiple faces but these faces were unfortunately false alarms by the face detector.

As to the first issue, in order to distinguish shots with a single face and shots with multiple faces, we can collect the following statistics of face information for each shot in a training data set:

- 1) the total number of faces in the shot
- 2) the number of frames having multiple faces within the shot
- 3) the total number of faces that labeled with high confidence by the face detector
- 4) the total number of faces where the face detector is only weakly confident.
- 5) the average number of faces within each image over the whole shot
- 6) the average number of highly confidently labeled faces within each image over the whole shot
- 7) the average number of faces labeled with weak confidence within each image over the whole shot
- 8) the total number of frames within the shot

As seen from the statistics (3), (4), (6) and (7), in addition to the counts of faces that are normally identified by the face detector, we also consider whether those faces are either highly confident or only weakly rated by the face detector. The reasoning is that the statistics of strongly believed and weakly believed faces may help distinguish between shots with a single face and shots with multiple faces.

For the second issue of mistakenly labeled faces, in order to distinguish the shots with true multiple faces from those with misidentified multiple faces, we collect the statistics of face scores and the face area determined by the face detector. Clearly, the intuition behind these features is that, when face scores are low, the chance for the face detector to misidentify a face will be higher. The same reasoning applies to the statistics of the face areas. There are totally five different statistics are collected in this category:

- 1) the average area of faces within a shot
- 2) the average face confidence scores within a shot
- 3) the standard deviation of the face scores within a shot
- 4) the maximum face score within a shot
- 5) the minimum face score within a shot

In order to show that these additional statistics may better distinguish between the 'people' category and the 'non-people' category, in Figure (2), we plot the distribution of shots with respect to the average confidence score for both categories.



Figure 2: The distribution of the percentage of shots with respect to the average face confidence score for both the 'people' category and the 'non-people' category.

As illustrated in Figure (2), the two distributions are quite different. Apparently, the 'people' category has a much higher spike around an average score equal to 2.4 than the category of 'non-people'. This phenomena indicates that, a shot with averaged face score around 2.4 is much more likely to be a shot of people than to be a shot of non-people.

2.2. Combination Model

Once we have obtained these statistics, the second step is to compute a combination model that is able to glue all the information together so that we can infer whether a shot belongs to the category of 'people' or not. In order to find the appropriate combination model, we need to consider the characteristics of this problem. First, as indicated in previous subsection, the features extracted from the face detector are strongly correlated with each other. For example, the averaged number of face for each image is strongly correlated with the averaged number of strongly believed faces for each image. Therefore, a good candidate of combination model should be able to take feature correlation into account. Secondly, as indicated from Figure (1) and (2), none of the individual extracted statistics is able to give a perfect distinction between the shots of people and shots of non-people. Therefore, a good candidate of combination model should be able to couple different pieces of evidence together.

Following to these two criteria, a simple classification model such as Naïve Bayes will definitely not be appropriate for this task because it usually assumes the features are independent. On the other hand, more complicated methods such as decision trees and support vector machines appear to fit this task better because they are usually robust to strong feature correlation and able to take into account the nonlinear coupling between features. Particularly, we examined three models for combination:

- Decision tree. The basic idea of a decision tree is to first find the most informative feature that is able to separate the shots of people from non-people and use that feature as the root node. Then, within each branch of the root node, the next most informative feature will be selected as the node on the second level of that branch. The tree will keep on growing until either all shots under all the leaf nodes are pure or a stop criterion is reached. More information about decision trees can be found in [1].
- 2) Support Vector Machine (SVM). The basic idea of a support vector machine is to find a decision boundary that can put shots of each category on different sides of the boundary. Furthermore, the decision boundary created by the support vector machine will be as far away from shots of both categories as possible. In another word, SVM is able to find the maximum margin between the two categories of shots. Lastly, a nonlinear coupling between different features is introduced through a nonlinear kernel function. More information about the support vector machine can be found in [2].
- 3) Logistic Regression Model. The logistic regression model assumes the prediction probability can be written as a linear exponential function of all the features and usually a maximum likelihood estimation is used for estimating parameters within the model. In general, the logistic regression model is good for combining correlated evidences for prediction. Similar to a support vector machine, the nonlinear coupling between features can be introduced through a nonlinear kernel function. More information about logistic regression models can be found in [3].

3. EXPERIMENT

In this section, we examine the effectiveness of our method. Particularly, we need to answer two questions:

- 1) Whether the extracted features are effective in distinguishing shots of 'people' from other shots?
- 2) Which combination model is best able to take all the features into account and give a better prediction?

3.1. Experimental Design

The datasets used for testing are the videos from the 'feature development collection' of the TREC2002 video track corpus [5], which together total 23 hours in length. The face detector used in this experiment was developed at CMU by Schneiderman [4], and applies statistical modeling to capture the variations in facial appearances. There were a total of 3840 shots within the test videos that contained a face. The baseline model used for comparison was to simply check if a shot contained images with multiple faces as described in the introduction section. The average classification error over a five-fold cross validation was used as the evaluation metric.

3.2. Experiment Results

The results for the baseline model together with the three combination models described in earlier sections are listed in Table 1.

 Table 1: Classification errors for the baseline model,
 decision tree,
 support vector machine and logistic
 regression

	Classification Error
Baseline	0.498
Decision Tree	0.409
Support Vector Machine	0.446
Logistic Regression	0.403

First, as indicated in Table 1, compared to the three combination models, the baseline model performs worst with classification error as 0.498. One big difference between the baseline model and three combination models is that the three combination models use all the extracted features while the baseline model only relies on a single statistics, namely the number of images with multiple faces within a shot. Therefore, the fact that all three combination models are able to outperform the baseline model indicates that the additional extracted features are useful for prediction. Secondly, within the three combination models, somewhat surprisingly, the support vector machine performs significantly worse than both the decision tree and the logistic regression model, with a classification error of 0.446. This fact suggests that, for the task of identifying the shots of 'people', the right choice of combination model can make a significant difference.

Finally, in the official TREC feature classification evaluation performed by assessors at NIST for finding shots with people, we use the set of extracted features described in section 2 and the decision tree as the combination model. Our system achieved the best performance of 9 submissions with a mean average precision of 27.1%. More details can be found in [6].

4. CONCLUSIONS

In conclusion, in order to find video shots with people using only face information, we need to extract more statistics beyond the simple count of the images with multiple faces within a shot. Furthermore, an appropriate choice of combination model can also be important in improving performance since many extracted statistics are strongly correlated. Of course, limiting our features to face information is probably the biggest problem in this approach to finding shots of 'people', particularly when the faces of the people may not be visible in the shot. More research work is needed to explore the use of body shapes and human motion for classifying shots with people..

6. REFERENCES

[1] R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA, 1993

[2] C. J. C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Knowledge Discovery and Data Mining, 2(2), 1998

[3] S. le Cessie and J.C. van Houwelingen, *Ridge Estimators in Logistic Regression*, Applied Statistics, Vol. 41, No. 1, pp. 191-201, 1997

[4] H. Schneiderman and T. Kanade, *Probabilistic Modeling of Local Appearance and Spatial Relationships of Object Recognition*, IEEE CVPR, Santa Barbara, 1998

[5] Web site for web retrieval in TREC, <u>http://www-nlpir.nist.gov/projects/trecvid</u>

[6] A.F.Smeaton and P. Over, (in press) The TREC 2002 Video Track Report, In E.M. Voorhees and D.K. Harman, The 11th Text Retrieval Conference (TREC 2002), Gaithersburg, MD, 2003.