PRE-ATTENTIONAL FILTERING IN COMPRESSED VIDEO

Juan M. Sánchez †, Ramon L. Felip ‡ and Xavier Binefa ‡

† Visual Century Research, SL.
Llacuna, 162. 08018, Barcelona, Spain.
‡ UPIIA, Dpt. d'Informatica.
Universitat Autonoma Barcelona. 08193, Bellaterra (Barcelona), Spain.

ABSTRACT

We propose the use of attentional cascades based on the DCT and motion information contained in an MPEG coded stream. An attentional cascade is a sequence of very efficient classifiers that reject a large number of negative candidate regions, while keeping all the positive candidates. Working directly on the compressed domain has two main advantages: computationally expensive features are already computed, and the stream is only partially decoded without the additional cost of full decompression, which will be reached by a very small number of the initial candidate regions. We have applied these concepts to skin color detection, as a pre-attentive filtering prior to face detection, and to text region detection with particular focus on license plates for vehicle identification. In both cases, a reduction of the number of candidate regions close to 95% is achieved, which turns into an enormous performance increase in video indexing processes.

1. INTRODUCTION

In general, object detection and recognition in video is a very costly task due to the large amount of information that has to be analyzed. Particularly, when no prior information about the possible location of the object is available, the whole image has to be searched by testing every sub-window through a classification scheme. However, an overwhelming majority of the tests are performed on negative sub-windows. This is the basic assumption for attentional cascades [1]. The main underlying idea is that very efficient classifiers can be designed to reject many of the negative sub-windows before more complex classifiers are called upon to achieve low false positive rates. In the case of large video databases, where thousands of hours of video have to be indexed and made available to the users, processing time is critical. In these databases, video is stored in compressed format, usually using one of the MPEG coding standards (MPEG 1, 2 or 4). The usual content indexing process requires decompression of the MPEG stream, computation of appropriate features and classification. However, MPEG coding standards are based on the computation of DCTs (frequential coding) and motion vectors (predictive coding) [2], which are readily available and provide information that can be used to focus attention on spatial and temporal regions of highest interest for indexing purposes. A survey of compressed-domain features used in audio and visual indexing and analysis of video by Wang et al. can be found in [3].

Following the concept of attentional cascades, very simple classifiers can be applied on the DCT coefficients and motion vectors of MPEG blocks and macro-blocks (MBs) to filter out negative ones very efficiently, before they are passed to more complex processing, probably through DCT inversion and full reconstruction of the images. The gain in processing speed is twofold: (1) the MPEG stream is only partially decoded, avoiding additional overhead, and (2) there is important information regarding motion, color and texture that is already computed.

2. SKIN COLOR DETECTION

Skin color detection is a very common filtering step prior to face detection. Only those sub-windows of the image where skin color has been detected will be analyzed by the face detector. Skin color detection can be performed directly on DCT coefficients. In an MPEG stream, the average color of a block is encoded in the DC coefficients of the Cb and Cr channels. This is particularly convenient, given that luminance variations are encoded in the Y component, which is not used in the representation of skin color. Chai and Bouzerdoum successfully tested in [4] skin color classification in the YCbCr color space using a Bayesian approach. The distribution of skin color in the Cb-Cr space is shown in fig. 1. We observe that it can be well approximated by a Gaussian distribution. Our filtering step consists of thresholding the likelihood of Cb-Cr test samples computed from

Work on MPEG at Visual Century Research is funded by CIDEM project R+DCOOP03 0029. Authors from the Universitat Autònoma de Barcelona are funded by CICYT grant TIC2003-06075.

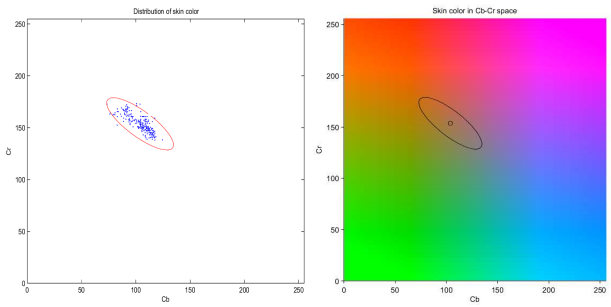

Fig. 1. Skin color distribution in the Cb-Cr space, and its approximation by a Gaussian. Left: skin color samples taken from real images and used to estimate the Gaussian parameters. Right: colors covered by this distribution, shown for a constant luminance ($Y = 128$).

the DC coefficients of each block. A threshold thr can be found such that 100% of the positive samples are not rejected. The implementation of this test can be very fast using a pre-computed 2D 256×256 look-up table (LUT), whose elements are:

$$LUT(i,j) = \begin{cases} 1 & \text{if } \mathcal{G}(i,j) > thr \\ 0 & \text{otherwise} \end{cases}$$

where $\mathcal{G}$ is the Gaussian distribution of skin colors in the Cb-Cr space. This LUT and threshold are graphically represented in fig. 1 (bottom).

We have tested this approach on 116 I-frames from a news video. With our filtering, 92.2% of the blocks were rejected. This means that only 7.8% of the total blocks will have to be decompressed and passed through more complex and slower face detection algorithms. We have also defined a second quick test in our cascade of classifiers, based on the aspect ratio of the bounding box around each connected group of skin color blocks. Let w and h be the width and height of the bounding box. We establish that, if the bounding box is around a face, the ratio w/h must be in the range $[0.35, 1]$. Also, the area of the bounding box $w \times h$ should be greater than 1, in order to avoid isolated false detections. With this second test, the percentage of rejected blocks is 94.42%, while 100% of the positive blocks are passed to subsequent classifiers. The number of positive blocks represents 5% of the original set. Only 0.58% of the blocks passed all the tests and were negative blocks. The process is depicted in fig. 2. Figure 3 shows examples on real images from our test set.

The method can be improved by increasing the set of skin color samples used for learning the Gaussian parameters, in order to account for natural variations of different skins. This variation would probably lead to a non-Gaussian distribution, so that other parametric distributions would be more appropriate. For instance, mixtures of Gaussians or other distributions could be used.

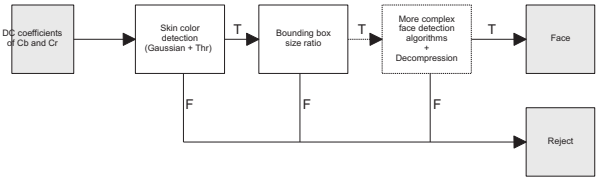

Fig. 2. Proposed attentional cascade for face detection using skin color filtering in the compressed domain. The decompression step is reached only by 5.58% of the blocks in our test set.

Fig. 3. Example of the proposed approach to skin color segmentation for face detection in the compressed domain. From left to right: original images, likelihood of each block w.r.t. the Gaussian skin color distribution in the Cb-Cr space (rescaled for proper visualization), results of skin color segmentation (step 1 of our attentional cascade), results after filtering bounding box size and aspect ratio (step 2 of the cascade).

3. LICENSE PLATE DETECTION

In the same way we have defined tests for quick rejection of negative blocks for skin color detection, we now define the tests that will form the first steps of an attentional cascade for license plate detection using compressed domain features. License plates are composed of type-written text. Text is characterized by high frequencies and high contrast. The DCT provides a frequential representation of the texture of a block, and is thus suitable for text detection. The DCT coefficients are sorted in zig-zag scan order, so that high frequency coefficients come last. Different tests can be perfomed in order to detect blocks with high frequency coefficients. For instance, we can identify the index of the last non-zero coefficient in zig-zag scan order. The higher this index is, the higher frequency exists in the block. However, text is usually defined by several different high frequencies. Therefore, a second approach is to count the number of non-zero coefficients in the block. The more non-zero coefficients, the more different frequencies.

Prior to text detection, we can introduce an optional motion detection step in the case of a static camera, so that license plates are only detected when they are on a moving vehicle and static MBs are rejected. Given a motion vec-

tor $\vec{m}$, a simple approach is to compute some motion feature $f(\vec{m})$ and apply a threshold t on it. In the presence of camera operation, this vector should be previously compensated for global motion using, for instance, the parametric estimation method described in [5], so that only foreground motion is detected. For simplicity, we show examples on static cameras, i.e. with no global motion.

A simple motion feature that can be computed is $f(\vec{m}) = |\vec{m}|$, i.e. the length of the vector. Some considerations have to be taken into account:

- There is no motion vector field for intra coded MBs, particularly in I pictures. In this case, interpolation of the motion measure can be considered in time, space, or both. A solution for this problem is out of the scope of this paper.
- The temporal distance between the current picture and the reference frame(s) is variable. Therefore, the motion measure must be normalized, so that it can be compared and thresholded properly. For a temporal distance of d units (frames), $\vec{m}/d$ is the average motion per time unit.
- Forward, backward or both motion vectors may be available. When both vectors are available, the average motion measure is considered:

$$f(\vec{m}) = \frac{1}{2}\left(\frac{|\vec{m}_f|}{d_f} + \frac{|\vec{m}_b|}{d_b}\right) \quad (1)$$

where $\vec{m}_f$ and $\vec{m}_b$ are the forward and backward motion vector, and d_f and d_b are the temporal distances from the current picture to the forward and backward reference frame, respectively.

Figure 4 shows an example of motion detection in two predictive frames of a sequence where a car is approaching a static camera.

Other quick tests can be defined in an attentional cascade in order to reject more negative candidates to license plate. For instance, we use again the bounding box aspect ratio. Tests can be defined also after decompression of candidate blocks. We have implemented a test that checks for the presence of typed characters in the candidate region. This is achieved by counting the number of oscillations of the projection profile. The most important fact is that only a very small percentage of the initial blocks, which includes all positive candidates, reach the decompression step. The rest of blocks are rejected by the first steps of the cascade. The proposed cascade is shown in fig. 5.

This approach has been tested on 133 I-frames from a sequence of different vehicles approaching a static camera. An example is shown in fig. 6. In this data set, the percentage of blocks rejected before the decompression step was 96.35%. After decompression, profile analysis raised the rejection rate up to 98.85%. The remaining 1.15% includes

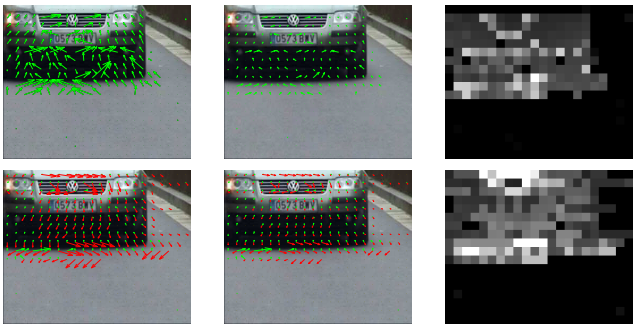

Fig. 4. Top: motion feature computed on motion vectors of a P picture, where only forward motion vectors are allowed. Bottom: feature computed on vectors of a B picture, where forward (green) and backward (red) motion vectors are allowed. Left: original vectors from the MPEG bitstream. Middle: vectors rescaled depending on the distance to the reference frame(s). The sequence of picture types in display order for the example bitstream is $IBBPBBP$.... Therefore, the distance d for a P picture is 3, while for a B picture we can have $d_f = 1$, $d_b = 2$ (case shown) or $d_f = 2$, $d_b = 1$. Right: motion feature computed for each MB and rescaled for better visualization. Whiter means faster motion.

all positive blocks, which represent 0.92% of the original set of blocks. Only 0.23% of the blocks passed all the tests and were negative blocks. We have also tested the method without any modification on a set of I-frames from a news video. Results are shown in fig. 7. In order to obtain accurate results in this other test set, the method should be adapted to the specifics of the domain. For instance, the aspect ratio filter would not be appropriate for generic text detection.

4. CONCLUSIONS

This paper has proposed the implementation of pre-attentive processes using MPEG compressed video. An attentional cascade composed of very simple classifiers allows us to quickly focus the attention on those macro-blocks that contain potentially interesting information, and only those will be fully decompressed and processed by more complex and slow algorithms to achieve even lower false positive rates. The classifiers are defined on MPEG pre-computed features, namely DCT coefficients and motion vectors, eliminating the time required for feature computation.

We have designed attentive cascades for skin color and text region detection, with particular interest on license plates. Both of them are pre-attentive processes largely used for video indexing, prior to face recognition and OCR. Skin color detection is based on the DC coefficients of the Cb and Cr color components and a LUT provides a very fast implementation of a Gaussian classifier. On the other hand, text

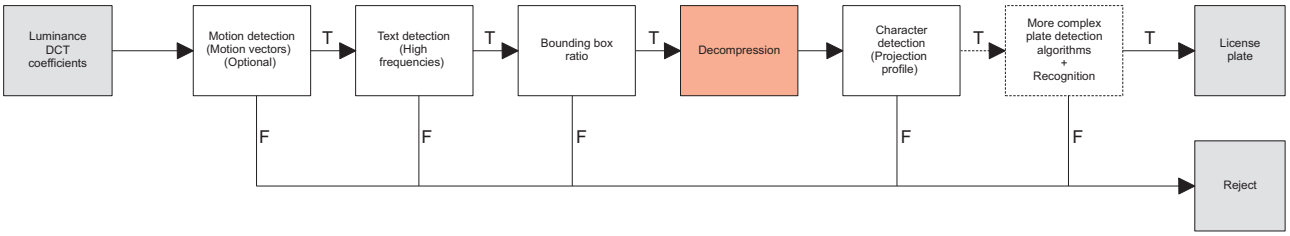

Fig. 5. Proposed attentional cascade for license plate detection. The decompression step is reached only by 3.65% of the blocks in our test set.

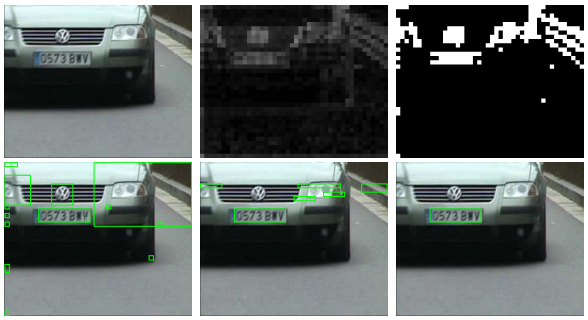

Fig. 6. Processing of an example frame by the attentional cascade. From left to right and top to bottom: original image, count of non-zero coefficients for each block, thresholded counts, results of high frequency block detection, results after bounding box aspect ratio filtering, results after profile analysis.

Fig. 7. Generic text detection in news videos using the same method of license plate detection.

region detection is based on the detection of high frequencies in the coefficients of the DCT. For the particular case of license plates, we have also applied a motion-based filter to detect plates on moving vehicles. In both cases, aspect ratio filtering helps to lower false positive rates. Faces and license plates are skin color and text regions, respectively, with very particular aspect ratio constraints.

The classifier cascades enormously reduce the number of MBs that will be processed by posterior recognition algorithms. In our tests, the cascades implemented reduced the number of candidate MBs down to 5%, which reverts into a reduction of nearly 95% in the time required to index contents. Results are summarized in tables 1 and 2.

	Step 0	Step 1	Step 2	Step 3
Faces	95%	2.8%	0.58%	N/A
License plates	99.08%	21.8%	2.5%	0.23%

Table 1. Summary of results: ratio of false positives in the complete test set (step 0), after DCT-based filtering (step 1), after aspect ratio filtering (step 2), and after profile projection based filtering (step 3, only for text detection). Decompression is done after step 2.

	Step 1	Step 2	Step 3
Faces	92.2%	94.42%	N/A
License plates	77.28%	96.35%	98.85%

Table 2. Summary of results: block rejection rates after each step in the cascade (see table 1). The decompression step is reached by around 5% of the blocks.

5. REFERENCES

[1] Paul Viola and Michael Jones, "Robust real-time object detection," in *Proc. Second International Workshop on Statistical and Computational Theories of Vision - Modeling, Learning, Computing, and Sampling*, July 2001.

[2] ISO/IEC, *14496-2: Information technology – Coding of audio-visual objects – Part 2: Visual*, December 2001.

[3] Hualu Wang, Ajay Divakaran, Anthony Vetro, Shih-Fu Chang, and Huifang Sun, "Survey of compressed-domain features used in audio-visual indexing and analysis," *Journal of Visual Communication and Image Representation*, vol. 14, pp. 150, June 2003.

[4] Douglas Chai and Abdesselam Bouzerdoum, "A bayesian approach to skin color classification in YCbCr color space," in *Proc. IEEE Region Ten Conference (TENCON'2000)*, Kuala Lumpur, Malaysia, September 2000, vol. II, pp. 421–424.

[5] Ramon Ll. Felip and Xavier Binefa, "Affine description of multiple motions in compressed domain," in *Proc. CCIA*, 2004.