Visual tools for assisting child pornography investigators

Pieter Eendebak[1], Wessel Kraaij[1], Stephan Raaijmakers[1], Elena Ranguelova[1],
Ork de Rooij[2], Andrew Thean[1], Marcel Worring[2]
[1]TNO, Delft, The Netherlands, [2] University of Amsterdam, The Netherlands
E-mail: [1]pieter.eendebak@tno.nl, [2]m.worring@uva.nl

Abstract: **In this paper we investigate the use of visual tools for efficient video search. The different techniques have been implemented in a demo called iDash and several of the techniques are actively used by the Dutch police force in the fight against child pornography.**

Keywords: video search, skin detection, video fingerprinting, object detection, semantic indexing , thread based video browsing

1 INTRODUCTION

With the arrival of the digital age, police in all countries dealing with child abuse and child pornography have found themselves facing new challenges. Broadband internet and affordable storage media have contributed to an explosion in the quantity of multimedia data confiscated by police as part of child pornography inquiries. In The Netherlands alone the amount of data confiscated by the police has increased from 4.5 to 152 terabytes between 2003 and 2006. The task of sifting through confiscated case material to identify child pornography can be tedious, time consuming, emotionally taxing and difficult. Increasingly, technologically-savvy pedophiles attempt to outwit law enforcement agencies by using technology, such as image-processing, video-editing and encryption software, to hide their tracks. The police have responded by taking the fight in the only direction possible, into the digital domain.

The main obstacle to be faced by the investigators is the huge amount of data that has to be inspected in order to find conclusive evidence for child abuse. This has to be achieved in the very limited time available to keep a person in custody without charge. In order to address these problems, suitable technological support is essential. During the last years TNO, the Dutch Forensic Institute (NFI), the University of Amsterdam and ZiuZ have worked together to develop tools for analyzing digital images and video.

Technology that has been developed so far consists of content-based image analysis tools (e.g. face detection, and skin or nudity detection), video fingerprinting and efficient video and image retrieval techniques that treat images analogical to text.

In this paper we describe the theoretical background and the practical benefits of these techniques and we propose their integration into a visual workbench called iDash tailored to forensic image analysis.

2 GENERIC OBJECT DETECTION

One topic of ongoing research is matching of specific objects and locations in large (usually unlabelled) image collections. A tool that allows large image databases to be visually 'googled' would help the police to link crimes by matching crime scenes. For example, a photo taken inside a suspect's home could be compared to child pornography databases. The key idea behind such a tool is to automatically detect *salient* (prominent, distinguishing) patches in an image and describe each patch independently to the viewing angle, distance to the camera and the acquisition conditions. In this way, finding the correspondences between two photos of the same scene is facilitated regardless of the transformations relating them.

A framework for generic object detection was pioneered by Schmid and Mohr [2] and can be summarized as: detect, describe, match. The three steps will be described in more detail below.

2.1 Detection: keypoints and salient regions

In many images there are regions which possess some distinguishing, invariant and stable property which can be detected independently with high repeatability. This property makes them a good choice for the representative image patches whose correspondence is sought. The detected salient regions should change *invariantly* with the transformation relating the two images [3,4]. The salient patches can be detected either as groups of image pixels in the vicinity of a *keypoint* or directly as *salient regions*.

Keypoints in images that can be detected repeatedly are mostly related to corner like structures. The Harris corner detector [1] (and variations) is a basic building block for many detectors. The Harris corner detector itself however is not scale or affine invariant. To introduce scale invariance often scale-spaces are introduced. Inspired by this, Lowe [6] proposed a method for extracting keypoints which are invariant to image scaling and rotation and partly invariant to

Corresponding author: Pieter Eendebak, TNO Science and Industry, Stieltjesweg 1, 2600 AD Delft, Phone: +31 15 269 28 53

change in illumination and camera viewpoint. This approach is known as the Scale Invariant Feature Transform (SIFT) as it transforms the image data into scale-invariant coordinates relative to local features. The SIFT features are the scale-space extrema, subject to a stability criterion (for details the reader is referred to [6]).

Although the SIFT algorithm performs very well it is not invariant to affine transformations. Since affine transformations appear with changes of the camera viewpoint several people have developed affine invariant detectors. A comparison of six state-of-the-art affine covariant region detectors is presented in [4]. For *structured* scenes, containing homogeneous regions with distinctive boundaries (as usually are the indoor scenes), the MSER (*Maximally Stable Extremal Region*) and IBR (*intensity-based region*) detectors perform best as they analyze the image isocontours directly.

Similarly to MSER, we proposed to analyze image isocontours by decomposing the image into binary cross-sections and computing two main types of saliency maps for each. The first type are the regions darker/brighter than their surroundings (similarly to MSER), and we propose a new type of salient regions manifested as significant irregularities on strong contrast borders. They are combined into a final map based on the stability of their support over the cross-sections.

Our detector uses morphological operators (for details the reader is referred to [7]), hence the name *Morphology-based Stable Salient Regions* (MSSR) detector. We have shown [7] that while the MSSR achieved comparable repeatability and matching performance to MSER and IBR, it is best in identifying perceptually salient regions. In Figure 1, bottom we have plotted all matched regions from the scene and the ones satisfying the spatial consistency contraints are shown connected with lines.

2.2 Region descriptors

In a second step of a generic matching application, the detected regions are encoded using a robust (invariant to geometric and photometric modifications) descriptor, and matching between the descriptors is performed. For the case of keypoints the descriptor is computed over the neighbourhood of the point, while in the case of the salient regions, the image values within the region (after normalisation) is used. A popular choice is the SIFT descriptor (usually of dimension 128) computed over the normalised regions- a 3D histogram of gradient location and orientations [6]. SIFT descriptors produce the best performance for different scene types, geometric and photometric transformations [5].

2.3 Matching

Descriptors are usually matched by using distance metrics (e.g. Euclidean or Mahalanobis distance) and selecting pairs with the shortest distance (nearest neighbour method). Since nearest neighbour queries in high dimensional spaces always have a worst-case quadratic running time, various approximations have been developed. Several geometric constraints can be added to further improve the matching. In Figure 1 we have plotted detected region, matches using a descriptor and the final matches after adding a spatial consistency constraint.

Figure 1: MSSR region detection. On top all detected regions, in the bottom corresponding regions and spatially consist matches

2.4 Visual words

Breaking down an image into an invariant set of image patches allows for applying insights from text retrieval. The image patches can be thought of as 'visual words' and a set of images can be treated as a set of 'documents' in which sets of visual words can be searched. The detected regions together with their descriptor are called visual words. Just as a normal text consists of words at specific locations, an image consists of visual words spread throughout this document. Using the analogy we can transfer search methods from text-retrieval into the visual domain.

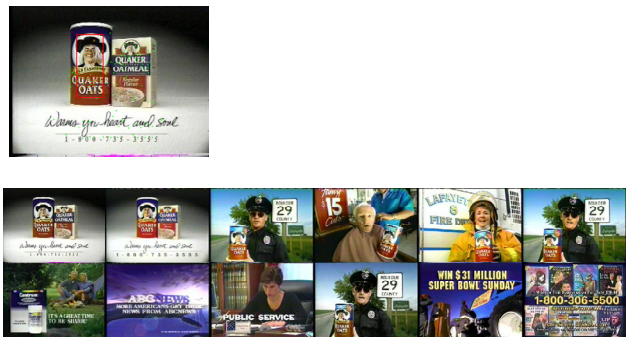

Figure 2: Query by example

In text searches several words are combined into groups. For example ‘color’, ‘colour’ and ‘colors’ are all combined into the same concept. In the visual domain this corresponds to grouping salient regions with similar descriptors into clusters.

Some of these clusters are not very useful for searching. For example when comparing English text the words `the' and `and' are usually omitted from the search. The same can be done by creating stoplists for visual words that are not very discriminative.

Using these techniques it is possible to search effectively through large image sets to locate an object of interest or a particular scene. In Figure 2 the result is shown of a query using an example image of a food box. The results contain not only the original image, but also many variations.

3 SKIN DETECTION

For the task of analyzing material containing (child) pornography a skin detector (or nudity detector) is very helpful. TNO and ZiuZ have developed a skin detector that is currently used in 70% of the Dutch police regions

3.1 Overview

Pixel-based skin detection involves the automated recognition of skin pixels in natural images, without any knowledge about texture, structure, or other types of aggregate information. The complexity of this problem arises partially from the wide color space of skin. Factors like illumination and color casts further add to the complexity of the problem. Usually, skin detection is performed in normalized color spaces. The choice for a suitable color space cannot be made independent of the type of classifier used. Choices here are mainly between parametric methods (such as Gaussian Mixture Models, or Maximum Entropy Classifiers ([14]) and non-parametric methods such as histogram-based Naïve Bayes classifiers. As noted by [11], the Naïve Bayes classifier is invariant with respect to color spaces, which is the reason why we adopted it in our work.

The Naïve Bayes skin classifier (or *skin probability map*) basically consists of a 3d-histogram and two priors, the probabilities of observing skin and non-skin. The 3d-histogram describes the chance of observing either skin or non-class given a certain RGB combination in the cells of the histogram. Every RGB combination in the training data is stored in a 3d-histogram. Apart from this, suitable smoothing techniques need to be applied to account for missing RGB combinations in the test data.

In order to cope with the problem of sparsity (unseen combinations of RGB values in the test data), the value range of a color dimension (red, green, blue) can be partitioned into a number of bins, after which the probabilities of observing skin or non-skin for a specific three-dimensional bin can be computed. For binning, several options are open, like fixed size binning, or minimum description based binning [12].

During testing, an image is again processed pixel by pixel. Every pixel is allocated to a certain 3d-bin, and the probability of skin for the bin the *RGB* combination is allocated to, is computed as follows.

$$P(skin \mid b(p_i)) = \frac{P(b(p_i) \mid skin)P(skin)}{P(b(p_i))}$$

$$P(b(p_i)) = P(b(p_i) \mid skin)P(skin) + P(b(p_i) \mid noskin)P(noskin)$$

The aggregate (non)skin probability for an entire image I, consisting of n pixels p (RGB combinations, with $b(p_i)$ denoting the bin this RGB combination is allocated to) is then simply

$$\overline{P}(skin \mid I) = \frac{1}{n}\sum_{i}^{n} P(skin \mid b(p_i))$$

$$\overline{P}(non-skin \mid I) = 1 - P(skin \mid I)$$

Alternatively, a pixel-wise Naïve Bayes classifier would consist of a simple thresholded decision rule mapping the bin for every pixel p_i in an image to a 0 (non-skin) or 1 (skin), which is in fact the classifier we used in our work. The threshold was estimated on heldout development data.

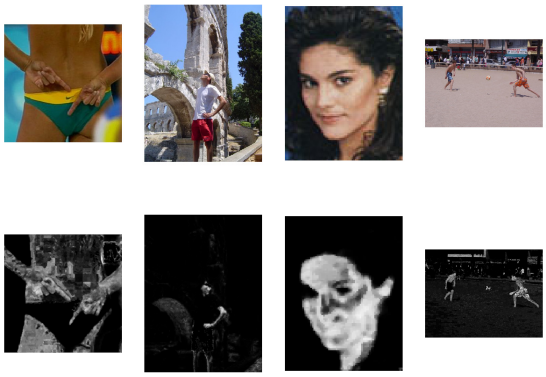

Figure 3: Skin detection applied to 4 random images

3.2 Experiments

In our experiments, we trained on a subset of 3000 images of the Compaq skin database [16]. This database consists of 4670 skin images and 8964 non-skin images supplied with manually crafted skin masks. These masks describe the partitioning of an image in skin and non-skin parts, from which the skin/non-skin probabilities can be estimated. We experimented with MDL based binning [12] but in the end opted for a simpler uniform binning in combination with smoothing. At the same with we experimented with pre-processing the images with a color correction algorithm.

For performance evaluation we consider the pixel classification performance and the image classification performance. For the pixel classification we simply count the number of pixels that are classified correct or incorrect as skin or non-skin. The results are presented as a ROC (Receiver Operator Curve) plot. As can be seen from Figure 4, binning

the skin data seems to consistently improve the performance a little bit, especially for higher true positive rates.

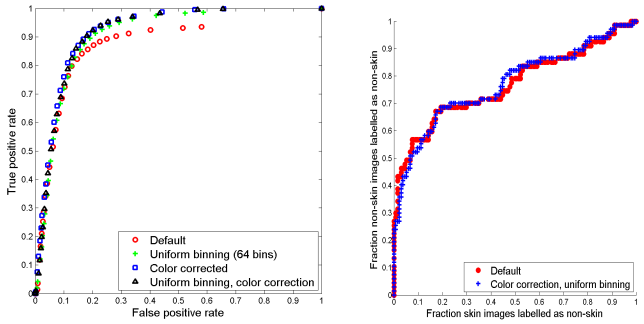

Figure 4: ROC for pixel classification (left) and entire image classification (right)

We implemented the skin filter in a prototypical image ranking system, ranking images on the basis of the aggregate pixel skin probabilities. Investigators can use this type of facility to focus on skin images first when browsing through a large repository of confiscated images, a process that benefits from high precision. If in the top *N* ranked images a (any) child pornography image is present, the suspect will be subject to further prosecution without further ado. Low precision would demote the workflow to manual inspection, in the worst case of all images. In Figure 5 the ROC curve is plotted for the skin detection algorithm applied to entire images, with or without color correction and uniform binning.

4 VIDEO FINGERPRINTING

As already noted the task of analyzing all confiscated material is a huge task. Over the years the police force has build up a database containing illegal material (both images and video).

New suspect material is first checked against already processed and categorized material with MD5 checksums. But often this test fails because of small transformations such as reencoding, logo insertion, lossy compression etc.

For identification purposes sometimes watermarking is used, i.e. the insertion of a hidden signal in the video. However insertion of the watermark has to be done by the producer of the material and this is not an option in the case of child pornography. Hence there arises the need for identification of videos based on the content of the video. These methods are generally called video fingerprinting.

Recently video printing technology has been under active development for digital rights management and copyright infringement applications [15]. The application of video fingerprinting to forensic research is quite new, although the first applications are appearing [13].

4.1 Overview

It is natural to make a distinction between CBCD (content-based copy detection, sometimes called near-copy detection) and CBVR (content-based video retrieval). With CBCD the goal is to find copies of videos (with small distortions). With CBVR the goal is to find videos with a (for humans) similar content. For forensic applications both applications are relevant and in practice video fingerprinting systems For example 2 videos of 2 different football matches should not be matched in CBCD, but they are both about football they can be matched within CBVR.

Figure 5: Similar videos retrieved by the TNO video fingerprinting system. Top match is CBCD, bottom match is CBVR.

4.2 Results

Since video fingerprinting is not new, the goal is not to develop completely new technology. Rather we want to identify the specific problems that arise in the context of child pornography.

One very specific problem is the huge legacy of video tapes (analog material) that is digitized for storage and transmission through the internet. Often this analog material has very specific distortions (noise, synchronization problems). Also during analogue to digital conversions it happens often that frames are missing from the data stream. This means that the video fingerprinting technique used must be robust to a certain percentage of missing frames. Related to this is the fact that child pornography material is often of relatively low quality as compared to commercial movies. Material is often created with inexpensive home cameras and distribution is done mainly over the internet where high compression is desirable. Further, quite often, different videos are compiled to new videos. This implies that fingerprinting techniques have to deploy local, time-stamped feature instead of global features (per video).

With these requirements in mind and the fact that video fingerprinting has to be applied to extremely large datasets, we have chosen for a technique that is easy to implement and is known to produce adequate performance. For local features

we use binary fingerprints derived from the individual frames. Each frame is divided into blocks and for each block the mean greyscale intensity is calculated, see Figure 6 below.

Figure 6: Haar based fingerprint

These binary fingerprints are indexed per video. Retrieval is done in several stages. From the new video material fingerprints are calculated. With these fingerprints and the index an initial search is performed. The result is a list of tentative matches. Each of these matches is then subjected to a more detailed comparison using fingerprints of both the database and the query video. Finally all the results are combined by removing duplicate results and combining overlapping results.

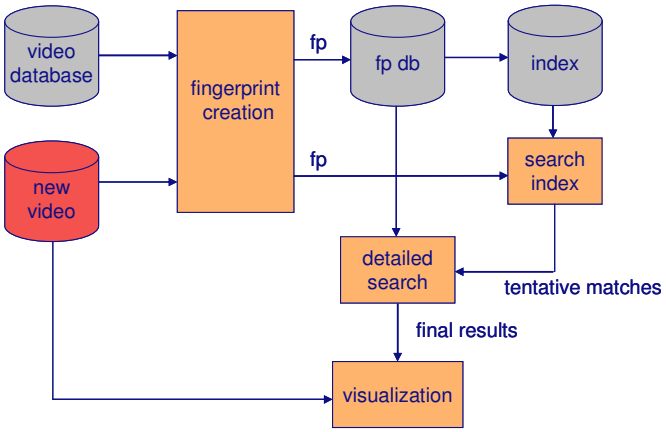

Figure 7: TNO video fingerprinting system

5 GENERIC SEMANTIC INDEXING

When looking at detectors for indexing child pornography data it is clear that some special detectors like the skin detector just described are important. However, when an investigator is searching through a large collection of videos to find any clue of victims or suspects a large set of concept detectors is useful. However, it is not feasible to develop specific detectors for each of them. Therefore, in addition to specific detectors we employ generic detector schemes.

The visual indexing process starts with computing a high-dimensional feature vector for each shot. In our system we use the Wiccest features as introduced in [10]. Wiccest features combine color invariance with natural image statistics. Color invariance aims to remove accidental lighting conditions, while natural image statistics efficiently represent image data.

Computing visual features and similarity is common practice in all interactive content based video retrieval systems. We now move on to the more specific topic of adding semantic indexing to the data, which is the process of associating every shot in the database with a measure of presence of the given concept. The central assumption in our semantic indexing architecture is that any video which is made with a purpose is the result of an authoring process. When we want to extract semantics from a digital broadcast video this authoring process needs to be reversed. For authoring-driven analysis we proposed the semantic pathfinder [9], composed of three analysis steps. It follows the reverse authoring process. Each analysis step in the path detects semantic concepts. In addition, one can exploit the output of an analysis step in the path as the input for the next one. The semantic pathfinder starts in the content analysis step. In this analysis step, we follow a data-driven approach, using both visual and textual information, of indexing semantics. The style analysis step is the second analysis step. Here we tackle the indexing problem by viewing a video from the perspective of production.

Finally, to enhance the indexes further, in the context analysis step, we view semantics in context. One would expect that some concepts, like vegetation, have their emphasis on content where the style (of the camera work that is) and context (of concepts like graphics) do not add much. In contrast, more complex events, like people walking, profit from incremental adaptation of the analysis to the intention of the author. The virtue of the semantic pathfinder is its ability to find the best path of analysis steps on a per-concept basis. The generic indexing structure has been used to create a lexicon of 101 concepts. Elements in the lexicon range from specific persons to generic classes of people, generic settings, specific and generic objects etc. See [9] for a complete list. The quality of the results varies widely. Some of the concepts have good accuracy, while some perform very poorly. There are two factors influencing the accuracy. The variety in appearance of the object and settings and whether the set of training samples supplied sufficiently covers this variety.

6 THE I-DASH APPLICATION

The different search technologies described in the previous sections have been integrated into a visual workbench called the Investigators Dashboard, or I-Dash for short. In this dashboard some additional novel techniques have been implemented which are described in the next paragraphs.

6.1 Thread based video browsing

All of the search techniques defined in the previous sections are based on rankings either by pre-ranking data based on semantic detectors, or example-driven: given some current focal shot, show near copies or video containing the same object. To allow for effective browsing it is advantageous to give the user the opportunity of browsing through all of these

dimensions. To do so we have introduced the notion of thread based browsing [8]. A thread is a linked sequence of shots in a specified order, based upon an aspect of their content. We define several thread types in our system. The most used form of threads is the query result thread: the result of a user constructed query. In this case the shots are linked because they all originate as results from the same query. Other forms of threads include visual threads, semantic threads, top-rank threads, textual threads and the time thread. The visual thread links shots together which share the same visual characteristics, so that shots next to each other are also visually similar. The semantic thread links shots together based on their detected semantic concept scores, so that shots next to each other both contain the same set of semantic concepts. The textual thread links shots to each other which contain the same ASR text. The time thread can be compared to the time line of a video. A special form of thread is the top-rank thread, which just connects the top N shots from every concept to each other; so that one thread length N is generated for every concept. The search engine supports two modes for displaying results. Both display modes show an active focal shot, and a collection of threads relevant to the focal shot. Both display modes use a fixed layout where the focal point is always the largest most centered shot on the screen, and all relevant threads are shown in a star formation around it. The user has only to choose between two actions: select, or bookmark, the current focal shot as a valid result, or switch focus to any of the neighbouring shots. As a third option the user can also use the mouse to directly bookmark any visible shot by clicking on it.

The CrossBrowser only displays the initial query results and the time thread, and thereby limits the user in the browsable dimensions. The RotorBrowser shows all relevant threads, including time, for each shot. The CrossBrowser allows movement through the initial query results, and for each result limited movement through the time thread. To preserve context the user is not allowed to leave the initial query results.

The RotorBrowser does allow the user to leave the initial query result set so the user can browse through anything that catches his interest. To prevent the user from “getting lost” a system of hotkeys was added to enable quick jumping back to the last initial query result. A screenshot of the CrossBrowser is shown in Figure 8.

6.2 Intelligent video playback

Playing a normal video search for events is not very efficient. For a video collection of 100 hours this would require an investigator to watch 100 hours of video. However, the human visual system is capable of processing much more data at the same time. In I-Dash video is played in six windows at the same time, at twice or more times the normal speed. In this way 1 hour of video can be reviewed in only 5 minutes or less.

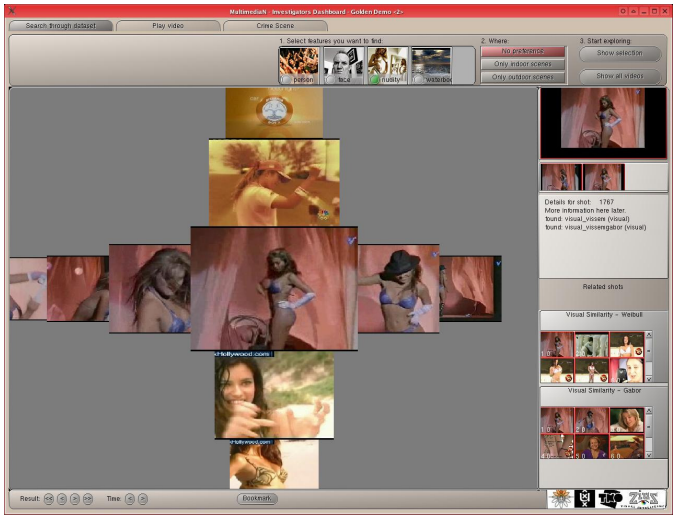

Figure 8: I-Dash with CrossBrowser

7 THE I-DASH PROJECT

The fight against production and distribution of child pornography by applying digital video analysis will be continued at the European level. The EU safer internet plus programme has funded the 2 year project IDASH with the following partners: University of Amsterdam, TNO, ZiuZ, University of Surrey, INESC and 5 EU police organisations as end users.

The project takes the I-dash visual workbench as a starting point. It will be installed at the start of the project at all end-user sites. In the mean time, applied R&D will adapt and improve the latest technology to this specific domain. The police will not only get a working solution right from the start, but competitive state-of-the-art methods in the course of the project as well.

The objectives are the development of an operational system capable of handling thousands of hours of videos potentially containing child pornography. Further, it will establish a European database with known child pornography and a standard for efficient data exchange between the various national police forces. The project will impact the European fight against child pornography by making investigators more efficient, and improving national and international collaboration.

8 CONCLUSIONS

Now that so many people have easy access to information and software the fight between law enforcement agencies and criminals is becoming more symmetrical. The latest research into image processing and information mining is crucial for keeping the police one step ahead of pedophiles. As new technology is employed in the fight against child pornography the difference between victory and defeat lies in deploying superior knowledge.

9 REFERENCES

[1] C. Harris and M. Stephens, A combined corner and edge detector, *Proceedings of the 4th Alvey Vision Conference*, pp. 147-151, 1988.

[2] C. Schmid and R. Mohr, Local grayvalue invariants for image retrieval, 1997.

[3] K. Mikolajcyzk and C. Schmid, Scale and Affine invariant interest point detectors, *Int. Journal of Computer Vision (IJCV)*, 1, vol.60, pp. 63-86, 2004

[4] K. Mikolajczyk et al., "A comparison of affine region detectors", *IJCV*, vol. 65, pp.43—72, 2005

[5] K.Mikolajczyk, C.Schmid, A performance evaluation of local descriptors, *PAMI* 27 (10), pp. 1615-1630, 2005

[6] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *IICV*, 60, 2, pp. 91-110, 2004

[7] Ranguelova, E., Pauwels, E. J. “Morphology-based Stable Salient Regions Detector”, *IVCNZ*, pp. 97-102, 2006

[8] O. de Rooij, C.G.M. Snoek, and M. Worring. Query on Demand Video Browsing. *Proc. of the ACM International Conference on Multimedia*, pp. 811-814, Augsburg, Germany, 2007.

[9] C.G. M. Snoek, M. Worring, J. Geusebroek, D.C. Koelma, F.J. Seinstra, and A.W. M. Smeulders. The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1678-1689

[10] J. van Gemert, J. Geusebroek, C. Veenman, C. Snoek, and A.W.Smeulders. Robust scene categorization by learning image statistics in context. *Proc. of CVPR Workshop on Semantic Learning Applications in Multimedia*, 2006.

[11] P. Kakumanu, S. Makrogiannis, N. Bourbakis, A survey of skin-color modeling and detection methods, *Pattern Recognition*, 40(3):1106–1122, 2007.

[12] P. Kontkanen and P. Myllymäki. Mdl histogram density estimation. In *International Conference on Artificial Intelligence and Statistics*, 2007.

[13] NSME - Google. http://www.missingkids.com/missingkids/servlet/NewsEventServlet?LanguageCountry=en_US&PageId=3644.

[14] B. Jedynak, H. Zheng, M. Daoudi, and D. Barret. Maximum entropy models for skin detection, 2002.

[15] A. Joly, C. Frelicot, and O. Buisson. Content-based video copy detection in large databases: a local fingerprints statistical similarity search approach, *Proc. IEEE International Conference on Image Processing 2005*, volume 1, pages I–505–8, 2005.

[16] M.J. Jones and J.M. Rehg. Statistical color models with application to skin detection, 2002.