Hindawi Publishing Corporation
EURASIP Journal on Image and Video Processing
Volume 2008, Article ID 195743, 10 pages
doi:10.1155/2008/195743

Research Article

Anthropocentric Video Segmentation for Lecture Webcasts

Gerald Friedland[1] and Raul Rojas[2]

[1] *International Computer Science Institute, 1947 Center Street, Suite 600 Berkeley, CA 94704-1198, USA*
[2] *Institut für Informatik, Freie Universität Berlin, Takustrasse 9, Berlin 14195, Germany*

Correspondence should be addressed to Gerald Friedland, fractor@icsi.berkeley.edu

Received 31 January 2007; Revised 16 July 2007; Accepted 12 December 2007

Recommended by Ioannis Pitas

Many lecture recording and presentation systems transmit slides or chalkboard content along with a small video of the instructor. As a result, two areas of the screen are competing for the viewer's attention, causing the widely known split-attention effect. Face and body gestures, such as pointing, do not appear in the context of the slides or the board. To eliminate this problem, this article proposes to extract the lecturer from the video stream and paste his or her image onto the board or slide image. As a result, the lecturer acting in front of the board or slides becomes the center of attention. The entire lecture presentation becomes more human-centered. This article presents both an analysis of the underlying psychological problems and an explanation of signal processing techniques that are applied in a concrete system. The presented algorithm is able to extract and overlay the lecturer online and in real time at full video resolution.

1. WEBCASTING CHALKBOARD LECTURES

If one wants to webcast a regular chalkboard presentation held in a classroom or lecture hall, there are mainly two ways to do this.

One possibility is to take a traditional video of the chalkboard together with the lecturer acting in front of it and then use standard webcasting products, such as Microsoft Windows Media or RealMedia to transmit the video into the Internet. The primary advantage of broadcasting a lecture this way is that the approach is rather straightforward: the setup for capturing a lecture is well known, and off-the-shelf Internet broadcasting software is ready to be used for digitizing, encoding, transmitting, and playing back the classroom event. Furthermore, the lecturer's workflow is not disturbed and nobody needs to become accustomed to any new devices. Even though some projects have tried to automate the process [1, 2], a major drawback of recording a lecture in the "conservative way" is that it requires additional manpower for camera and audio device operation. Yet the video compression techniques used by traditional video codecs are not suitable for chalkboard lectures: video codecs mostly assume that higher frequency features of images are less relevant. This produces either an unreadable blurring of the board handwriting or a bad compression ratio. Vector format storage is not only smaller, it is also favorable because semantics is preserved. After a lecture has been converted to video, it is, for example, not possible to delete individual strokes or to insert a scroll event without recalculating and rendering huge parts of the video again. Some projects have therefore tried to recognize board content automatically; see, for example, [3]. In most cases, however, this is hard to achieve, because chalkboard drawings are sometimes also difficult to read due to their low contrast. Figure 1 shows an example of a traditional chalkboard lecture webcast with a commercial Internet broadcasting program.

Knowing the disadvantages of the conservative approach, several researchers have investigated the use of pen-based computing devices, such as interactive whiteboards or tablet PCs to perform lecture webcasting (see, e.g., [4, 5]). Using a pen-based device provides an interesting alternative because it captures handwriting and allows storage of the strokes in a vector-based format. Vector-based information requires less bandwidth, can be transmitted without loss of semantics, and is easily rendered as a crisp image on a remote computer. Still, a disadvantage is the low resolution of these devices and the requirement for professors to change some teaching habits and technical accessories. One of the systems that supports the creation of remote lectures held using a pen-input device is our E-Chalk system [6], created in 2001. E-Chalk

Figure 1: A chalkboard lecture captured and replayed with commercial Internet broadcasting systems. Due to the lossy compression and the low contrast, the chalkboard content is difficult to read.

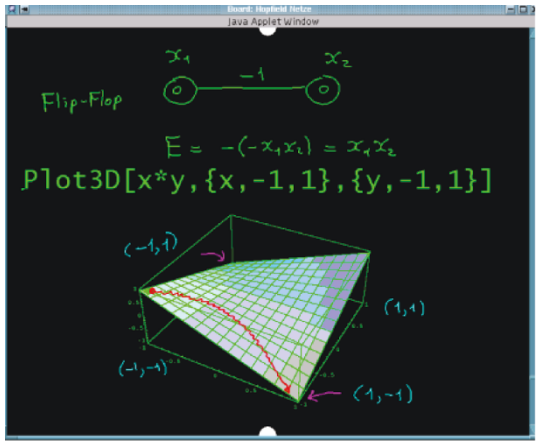


Figure 2: A transmission of chalkboard lecture held with an interactive whiteboard. The creation of the board content is transmitted as vector graphics while the voice of the instructor is played back in the background. This way of doing remote lecturing is bandwidth efficient and effective but lacks the perception of personality.

records the creation of the board content together with the audio track of the lecturer and transmits both synchronized over the Internet. The lecture can be received remotely either using a Java applet client or using MPEG-4 (see Figure 2).

2. HANDWRITING ONLY IS NOT SATISFYING

During an evaluation, many students reported they found it disturbing that the handwritten objects on the board appear from the void during distance replay. The lecture appears impersonal because there is no person acting in front of the board. The replay lacks important information because the so-called "chalk and talk" lecture actually consists of more

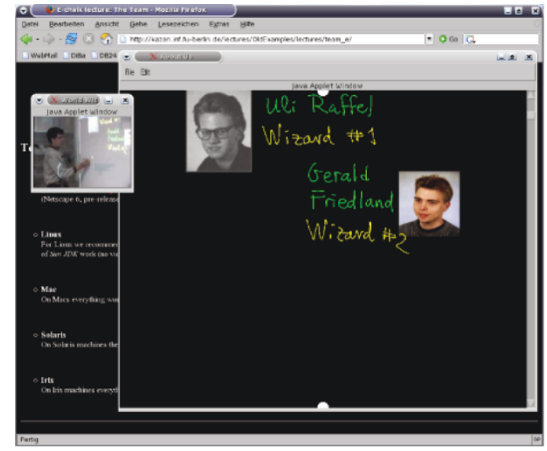


Figure 3: An example of the use of an additional video client to convey an impression of the classroom context and a view of the instructor to the remote student. This simple side-by-side replay of the two visual elements results in technical problems and is cognitively suboptimal.

than the content of the board and the voice of the instructor. Often, facial expressions of the lecturer bespeak facts beyond verbal communication and the instructor uses gestures to point to certain facts drawn on the board. Sometimes, it is also interesting to get an impression of the classroom or lecture hall. Psychology suggests (see, e.g., [7]) that face and body gestures contribute significantly to the expressiveness of human communication. The understanding of words partly depends on gestures as they are also used to interpret and disambiguate the spoken word [8]. All these shortcomings are aggravated by board activity being temporarily abandoned for pure verbal explanations or even nonverbal communication. In order to transport this additional information to a remote computer, we added another video server to the E-Chalk system. As shown in Figure 3, the video pops up as a small window during lecture replay. The importance of the additional video is also supported by the fact that several other lecture-recording systems (compare, e.g., [9]) have also implemented this functionality and the use of an additional instructor or classroom video is also widely discussed in empirical studies. Not only does an additional video provide nonverbal cues on the confidence of the speaker at certain points—such as moments of irony [10]—several experimental studies (for an overview, refer to [11]) have also provided evidence that showing the lecturer's gestures has a positive effect on learning. For example, [12] has reported that students are better motivated when watching lecture recordings with slides and video in contrast to watching a replay that only contains slides and audio. Reference [13] also shows in a comparative study that students usually prefer lecture recordings with video images over those without.

3. SPLIT ATTENTION

The video of the instructor conveys nonverbal information that several empirical studies have shown to be of value

Figure 4: The approach presented in this article (images created using the algorithm presented in this article). The remote listener watches a remote lecture where the extracted lecturer is overlaid semitransparently onto the dynamic board strokes which are stored as vector graphics. Upper row: original video; second row: segmented lecturer video; third row: board data as vector graphics. In the final fourth row, the lecturer is pasted semitransparently on the chalkboard and played back as MPEG-4 video.

for the student. There are, however, several reasons against showing a video of the lecturer next to slides or the blackboard visualization. The video shows the instructor together with the board content; in other words, the board content is actually transmitted redundantly. On low-resolution devices, the main concern is that the instructor video takes up a significant amount of space. The bigger the video is, the better nonverbal information can be transmitted. Ultimately, the video must be of the size of the board to convey every bit of information. As the board resolution increases because electronic chalkboards become better, it is less and less possible to transmit the video side-by-side with the chalkboard content. Even though there still might be solutions for these layout issues, a more heavily discussed topic is the issue of *split attention*.

In a typical E-Chalk lecture with instructor video, there are two areas of the screen competing for the viewer's eye: the video window showing the instructor, and the board or slides window. Several practical experiments that are related to the work presented here have been described in [14, 15]. Glowalla [13] tracked the eye movements of students while watching a lecture recording that contains slides and an instructor video. His measurements show that students spend about 70 percent of the time watching the instructor video and only about 20 percent of the time watching the slides. The remaining 10 percent of the eye focus was lost for activities unrelated to lecture content. When the lecture replay only consists of slides and audio, students spend about 60 percent of the time looking at the slide. Of course, there is no other spot to focus attention on in the lecture recording. The remaining 40 percent, however, were lost in distraction. The results are not directly transferable to electronic chalkboard-based lecture replays because the slides consist of static images and the chalkboard window shows a dynamic replay [16]. However, motion is known to attract human attention more than static data (see, e.g., [17]), it is therefore likely that the eyes of the viewer will focus more often on the chalkboard, even when a video is presented. Although the applicability of Glowalla's study to chalkboard lectures is yet to be proven, the example shows that, on a typical computer screen, two areas of the screen may be competing well for attention. Furthermore, it makes sense to assume that alternating between different visual attractors causes cognitive overhead. Reference [18] already discussed this issue and provided evidence that "Students presented a split source of information will need to expend a portion of their cognitive resources mentally integrating the different sources of information. This reduces the cognitive resources available for learning."

4. A SOLUTION

Concluding what has been said in the last two sections, the following statements seem to hold.

(i) Replaying a traditional video of the (electronic) chalkboard lecture instead of using a vector-based representation is bandwidth inefficient, visually suboptimal, and results in a loss of semantics.
(ii) If bandwidth is not a bottleneck, showing a video of the instructor conveys valuable nonverbal content that has a positive effect on the learner.
(iii) Replaying such a video in a separate window side-by-side with the chalkboard content is suboptimal because of layout constraints and known cognitive issues.

The statements lead to an enhanced solution for the transmission of the nonverbal communication of the instructor in relation to the electronic chalkboard content. The instructor is filmed as he or she acts in front of the board by using a standard video camera and is then separated by a novel video segmentation approach that is discussed in the forthcoming sections. The image of the instructor can then be overlaid on the board, creating the impression that the lecturer is working directly on the screen of the remote student. Figure 4 shows the approach. Face and body gestures of the instructor then appear in direct correspondence to the board events. The superimposed lecturer helps the student to better associate the lecturer's gestures with the board content. Pasting the instructor on the board also reduces bandwidth and resolution requirements. Moreover, the image of the lecturer can be made opaque or semitransparent. This enables the student to look through the lecturer. In the digital world, the instructor does not occlude any board content, even if he or she is standing right in front of it. In other words, the digitalization of the lecture scenario solves another "layout" problem that occurs in the real world (where it is actually impossible to solve).

5. RELATED APPROACHES

5.1. *Transmission of gestures*

The importance of transmitting gestures and facial expressions is not specific to remote chalkboard lecturing. In

a computer-supported collaborative work scenario, people first work together on a drawing and then want to discuss it by pointing to specific details of the sketch. For this reason, several projects have begun to develop means to present gestures in their corresponding context.

Two early projects of this kind were called *Video-Draw* [19] and *Video Whiteboard* [20]. On each side, a person can draw atop a monitor using whiteboard pens. The drawings together with the arms of the drawer were captured using an analog camera and transmitted to the other side, so that each side sees the picture of the remote monitor overlaid on their own drawings. Polarizing filters were used to omit video feedback. The VideoWhiteboard uses the same idea, but people are able to work on a large upright frosted glass screen and a projector is used to display the remote view. Both projects are based on analog technology without any involvement of the computer.

Modern approaches include a solution by [21] that uses chroma keying for segmenting the hands of the acting person and then overlaying it ona shared drawing workspace. In order to use chroma keying, people have to gesture atop a solid-blue surface and not on top of their drawing. This is reported to produce confusion in several situations. *LIDS* [22] captures the image of a person working in front of a shared display with a digital camera. The image is then transformed via a rough background subtraction into a frame containing the whiteboard strokes and a digital shadow of the person (in gray color). The *VideoArms* project by [23] works with touch-sensitive surfaces and a web camera. After a short calibration, the software extracts skin colors and overlays the extracted pixels semitransparently over the image of the display. This combined picture is then transmitted live to remote locations. The system allows multiparty communication, that is, more than two parties. Reference [24] presents an evaluation of the VideoArms project along with LIDS. He argues that the key problem is still a technical one: "VideoArms" images were not clear and crisp enough for participants. [...] the color segmentation technique used was not perfect,producing on-screen artifacts or holes and sometimes confusing users."

In summary, the presented approaches tried to work around either object extraction or the technical requirements for the segmentation that made the systems suboptimal. It is therefore important that the lecturer segmentation approach is either easily used in classroom and/or after a session; and technical requirements do not disturb the classroom lecture.

5.2. Segmentation approaches

The standard technologies for overlaying foreground objects onto a given background are chroma keying (see, e.g., [25]) and background subtraction (see, e.g., [26]). For chroma keying, an actor is filmed in front of a blue or green screen. The image is then processed by analog devices or a computer so that all blue or green pixels are set to transparent. Background subtraction works similarly: a static scene is filmed without actors once for calibration. Then, the actors play normally in front of the static scene. The filmed images are then subtracted pixel by pixel from the initially calibrated scene. In the output image, regions with pixel differences near zero are defined transparent. In order to suppress noise, illumination changes, reflections of shadows, and other unwanted artifacts, several techniques have been proposed that extend the basic background subtraction approaches. Mainly, abstractions are used that substitute the pixelwise subtraction by using a classifier (see, e.g., [27]). Although nonparametric approaches exist, such as [28], per-pixel Gaussian Mixture Models (GMM) are the standard tools for modeling a relatively static background (see, e.g., [29]). These techniques are not applicable to the given lecturer segmentation problem because the background of the scene is neither monochromatic nor fixed. During a lecture, the instructor works on the electronic chalkboard and thus causes a steady change of the "background."

Even though the background color of the board is black (RGB value (0,0,0)), the camera sees a quite different picture. In particular, noise and reflections make it impossible to threshold a certain color. Furthermore, while the instructor is working on the board, strokes and other objects appear in a different color from the board background color so that several colors have to be subtracted. A next experiment consisted of matching the blackboard image on the screen with the picture seen by the camera and subtracting them. During the lecture recording, an additional program regularly makes screenshots. The screenshots contained the board content as well as any frame insets and dialogs shown on the screen. However, subtracting the screenshots from the camera view was impractical. In order to match the screen picture and the camera view, lens distortion and other geometric displacements have to be removed. This requires a calibration of the camera before each lecture. Taking screenshots with a resolution of 1024×768 pixels or higher is not possible at high frame rates. In our experiments, we were able to capture about one screenshot every second and this took almost a hundred percent of the CPU time. Furthermore, it is almost impossible to synchronize screen grabbing with the camera pictures. In a regular lecture, many things may happen during a second. Still, a matching between the colors in the camera view and the screenshots has to be found.

Much work has been done on tracking (i.e., localization) of objects for computer vision, for example, in robotic soccer [30], surveillance tasks [31], or traffic applications [32]. Most of these approaches concentrate on special features of the foreground, and in these domains, real-time performance is more relevant than segmentation accuracy as long as the important features can be extracted from each video frame. Separating the foreground from more or less dynamic background is the object of current research.

Many systems use complex statistical methods that require intensive calculations not possible in real time (e.g., [33]) or use domain-specific assumptions (a typical example is [34]). Numerous computationally intensive segmentation algorithms have also been developed in the MPEG-4 research community, for example, [35]. For the task investigated here, the segmentation should be as accurate as possible. A real-time solution is needed for live transmission of lectures. Reference [36] presents a video segmentation approach that uses the optical flow to discriminate between layers of moving

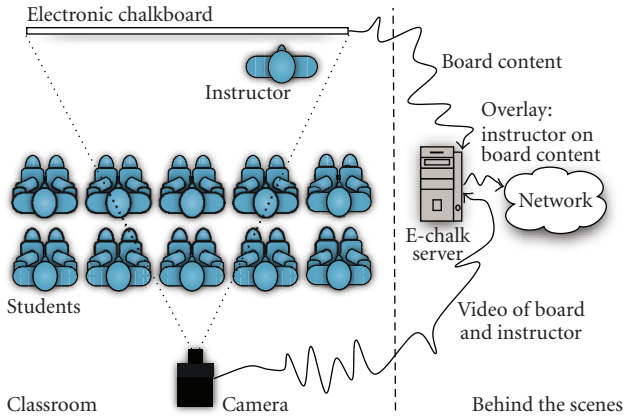


FIGURE 5: A sketch of the setup for lecturer segmentation. An electronic chalkboard is used to capture the board content and a camera records the instructor acting in front of the board.

pixels on the basis of their direction of movement. In order to be able to track an object, the algorithm has to classify it as one layer. However, a set of pixels is grouped into a layer if they perform the same correlating movement. This makes it a useful approach for motion-based video compression but it is not perfectly suited for object extraction. Reference [37] is combining motion estimation and segmentation by intensity through a Bayesian belief network to a spatiotemporal segmentation. The result is modeled in a Markov-Random field, which is iteratively optimized to maximize a conditional probability function. The approach relies purely on intensity and movement, and is therefore capable of segmenting grey scale. Since the approach also groups the objects by the similarity of the movement, the same limitations as in [36] apply. No details on the real time capability were given.

6. SETUP

In E-Chalk, the principal scenario is that of an instructor using an electronic chalkboard in front of the classroom. The camera records the instructor acting in front of the board such that exactly the screen showing the board content is recorded. With a zoom camera, this is easily possible from a nondisturbing distance (e.g., from the rear of the classroom); and lens distortion is negligible. In this article, it is assumed that the instructor operates using an electronic chalkboard with a rear projection (e.g., a StarBoard) rather than one with front projection. The reason for this is that when a person acts in front of the board and a front projector is used, the board content is also projected onto the person. This makes segmentation very difficult. Furthermore, given a segmentation, the projected board artifacts disturb the appearance of the lecturer. Once set up, the camera does not require operation by a camera person. In order to ease segmentation, light changes and (automatic) camera adjustments should be inhibited as much as possible. Figure 5 shows a sketch of the setup.

7. INSTRUCTOR EXTRACTION

A robust segmentation between instructor and background is hard to find using motion statistics. However, getting a subset of the background by looking at a short series of frames is possible. Given a subset of the background, the problem reduces to classifying the rest of the pixels as to either belonging to the background or not. The idea behind the approach presented here is based on the notion of a color signature. A color signature models an image or part of an image by its representative colors. This abstraction technique is frequently used in different variants in image retrieval applications, where color signatures are used to compare patterns representing images (see, e.g., [38, 39]). A variation of the notion of a color signature is able to solve the lecturer extraction problem and is useful for a variety of other image and video segmentation tasks. Further details on the following algorithm are available in [40, 41]. The approach presented here is based on the following assumptions. The hardware is set up as described in Section 6; the colors of the instructor image are overall different from those in the rest of the image; and during the first few seconds after the start of the recording, there is only one instructor and he or she moves in front of the camera. The input is a sequence of digitized YUV or RGB video frames, either from a recorded video or directly from a camera. The following steps are performed.

(1) Convert the pixels of each video frame to the CIELAB color space.
(2) Gather samples of the background colors using motion statistics.
(3) Find the representative colors of the background (i.e., build a color signature of the background).
(4) Classify each pixel of a frame by measuring the distance to the color signature.
(5) Apply some postprocessing steps, for example, noise reduction, and biggest component search.
(6) Suppress recently drawn board strokes.

The segmented instructor is then saved into MPEG-4 format. The client scales the video up to board size and replays it semitransparently.

7.1. Conversion to CIELAB

The first step of the algorithm is to convert each frame to the CIELAB color space [42]. Using a large amount of measurements (see [43]), this color space was explicitly designed as a perceptually uniform color space. It is based on the opponent-color theory of color vision [44, 45]. The theory assumes that two colors cannot be both green and red or blue and yellow at the same time. As a result, single values can be used to describe the red/green and the yellow/blue attributes. When a color is expressed in CIELAB, L defines lightness, a denotes the red/green value, and b the yellow/blue value. In the algorithm described here, the standard observer and the D65 reference white [46] are used as an approximation to all possible color and lighting conditions that might appear in an image. CIELAB is still not the optimal perceptual color space (see, e.g., [47]) and the aforementioned assumption

sometimes leads to problems. But in practice, the Euclidean distance between two colors in this space better approximates a perceptually uniform measure for color differences than in any other color space, like YUV, HSI, or RGB.

7.2. Gathering background samples

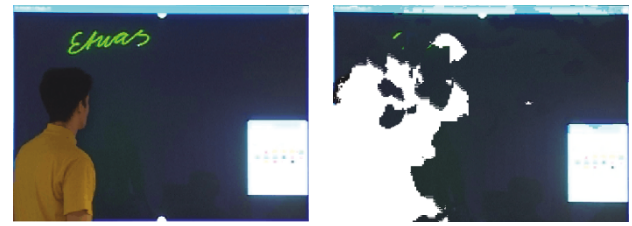
It is hard to get a background image for direct subtraction. The instructor can paste images or even animations onto the board; and when the instructor scrolls a page of board content upwards, the entire screen is updated. However, the instructor sometimes stands still producing fewer changes than the background noise. The idea is thus to extract only a representative subset of the background that does not contain any foreground for further processing.

To distinguish noise from real movements, we use the following simple but general model. Given two measurements m_1 and m_2 of the same object, with each measurement having a maximum deviation e from the real world due to noise or other factors, it is clear that the maximum possible deviation between m_1 and m_2 is $2e$. Given several consecutive frames, we estimate e to find out which pixels changed due to noise and which pixels changed due to real movement. To achieve this, we record the color changes of each pixel (x, y) over a certain number of frames $t(x, y)$, called the recording period. We assume that in this interval, the minimal change is caused only by noise. The image data is continuously evaluated. The frame is divided into 16 equally sized regions and changes are accumulated in each region. Under the assumption that at least one of these regions was not touched by any foreground object (the instructor is unlikely to cover the entire camera region), $2e$ is estimated to be the maximum variation of the region with the minimal sum. We then join all pixels of the current frame with the background sample that during the recording period $t(x, y)$ did not change more than our estimated $2e$. The recording period $t(x, y)$ is initialized within one second and is continuously increased for pixels that are seldom classified as background. This is done to avoid adding a still-standing foreground object to the background buffer. In our experiments, it took a few seconds for enough pixels to be collected to form a representative subset of the background. We call this time period the initialization phase. The background sample buffer is organized as an aging FIFO queue. Figure 6 shows typical background samples after the initialization phase.

The background sample is fed into the clustering method described in the next section. Once built up, the clustering is only updated when more than a quarter of the underlying background sample has changed. However, a constant updating is still needed in order to be able to react to changing lighting conditions.

7.3. Building a model of the background

The idea behind color signatures is to provide a means for abstraction that sorts out individual outliers caused by noise and small error. A color signature is a set of representative colors, not necessarily a subset of the input colors. While the set of background samples from Section 7.2 typically consists



(a) (b)

Figure 6: Using motion statistics, a sample of the background is gathered. The images show the original video (a) and known background that was reconstructed over several frames (b). The white regions constitute the unknown region.

of a few hundred thousand colors, the following clustering reduces the background sample to its representative colors, usually about a few hundred. The known background sample is clustered into equally sized clusters because in CIELAB space specifying a cluster size means specifying a certain perceptual accuracy. To do this efficiently, we use the modified two-stage k-d tree [48] algorithm described in [49], where the splitting rule is to simply divide the given interval into two equally sized subintervals (instead of splitting the sample set at its median). In the first phase, approximate clusters are found by building up the tree and stopping when an interval at a node has become smaller than the allowed cluster diameter. At this point, clusters may be split into several nodes. In the second stage of the algorithm, nodes that belong to several clusters are recombined. To do this, another k-d tree clustering is performed using just the cluster centroids from the first phase. We use different cluster sizes for L, a, and b axes. The values can be set by the user according to the perceived color diversity on each of the axes. The default is 0.64 for L, 1.28 for a, and 2.56 for the b axis. For efficiency reasons and for further abstraction, clusters that contain fewer than 0.1% of the pixels of the entire background sample are removed. The constants were learned with a set of benchmark images using a genetic algorithm.

The k-d tree is explicitly built and the interval boundaries are stored in the nodes. Given a certain pixel, all that has to be done is to traverse the tree to find out whether it belongs to one of the known background clusters or not. Figure 7 shows an example color signature.

7.4. Postprocessing

The pure foreground/background classification based on the color signature will usually select some individual pixels in the background with a foreground color and vice versa, resulting in tiny holes in the foreground object. The wrongly classified background pixels are eliminated by a standard *erode* filter operation while the tiny holes are filled by a standard *dilate* operation. A standard Gaussian noise filter smoothing reduces the number of jagged edges and hard corners. A biggest connected component search is then performed. The biggest connected component is considered to be the instructor, and all other connected components

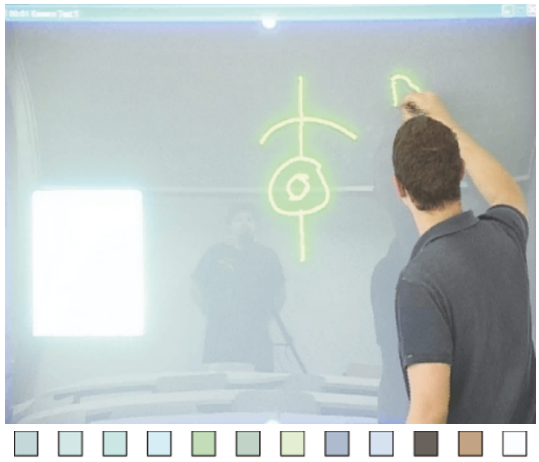
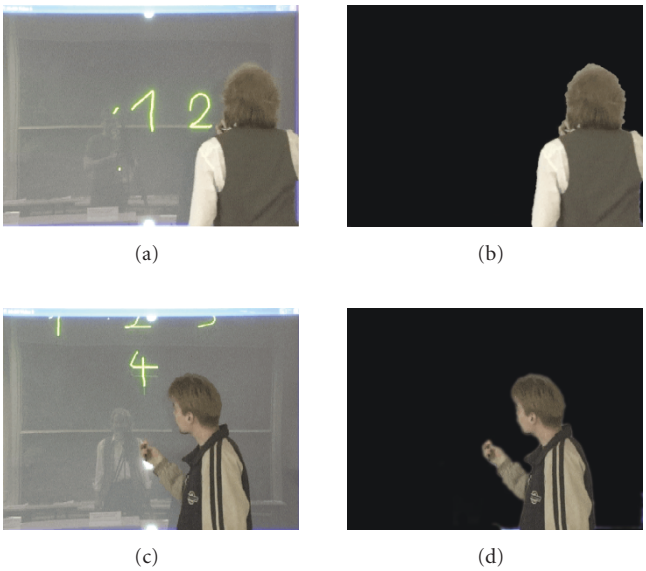


Figure 7: Original picture (above) and a corresponding color signature representing the entire image (below). For visualization purposes, the color signature was generated using very rough limits so that it contains only a few representative colors.



(a) (b) (c) (d)

Figure 8: Two examples of color-segmented instructors. Original frames are shown on the left, segmented frames are shown on the right. The frame below shows an instructor scrolling the board, which requires an update of many background samples.

(mostly noise and other moving or newly introduced objects) are eliminated from the output image. Figure 8 shows two sample frames of a video where the instructor has been extracted as described here.

7.5. Board stroke suppression

As described in Section 7.2, the background model is built using statistics over several frames. Recently inserted board content is therefore not part of it. For example, when an animation is used on the board, a huge amount of new board content is shown on the board in a short time. With the connected component analysis performed for the pixels classified as foreground, most of the unconnected strokes and other blackboard content have already been eliminated. In order to suppress strokes just drawn by the lecturer, all colors from the board system's color palette are inserted as cluster centroids to the k-d tree. However, as the real appearance of the writing varies with both projection screen and camera settings and with illumination, not all of the board activities can be suppressed. Additionally, strokes are surrounded by regions of noise that make them appear to be foreground. In order to suppress most of those thinner objects, that is, objects that only expand a few pixels in the X and/or the Y-dimensions are eliminated using an erode operation. Fortunately, a few remaining board strokes are not very disturbing because the segmented video is later overlaid on the board drawings anyways. Figure 9 compares two segmented frames with and without board stroke suppression.

(a) (b)

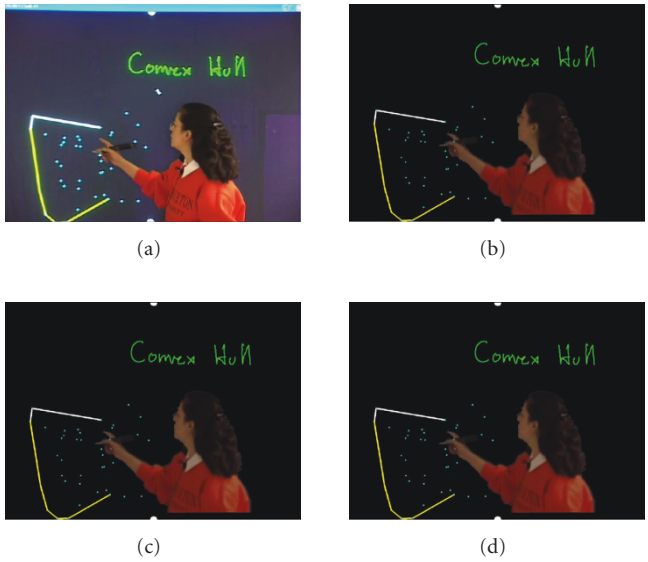
Figure 9: Board drawings that are connected to the instructor are often considered foreground by the classification. An additional board stroke suppression eliminates these artifacts. (a): the result of the color signature classification. (b): after applying a postprocessing step to eliminate board strokes.

8. LIMITS OF THE APPROACH

The most critical drawback of the presented approach is color dependence. Although the instructor videos are mostly well separable by color, the approach fails when parts of the instructor are very similar to the background. When the instructor wears a white shirt, for example, the segmentation sometimes fails because dialog boxes often also appear as white to the camera.

The presented approach requires that the instructor moves at least during the initialization phase. During our experimental recordings, we did not find this to be impractical. However, it requires some knowledge and is therefore prone to usage errors. The quality of the segmentation is suboptimal if the instructor does not appear in the picture during the first few frames or does not move at all.

Another problem is that if the instructor points at a rapidly changing object (e.g., an animation on the board screen) of a similar color structure, the instructor and the animation might both be classified as foreground. If they are connected somehow, the two corresponding components could be displayed as the single biggest component.



(a) (b)

(c) (d)

Figure 10: The final result: the instructor is extracted from the original video (left) and pasted semitransparently over the vector-based board content (right).

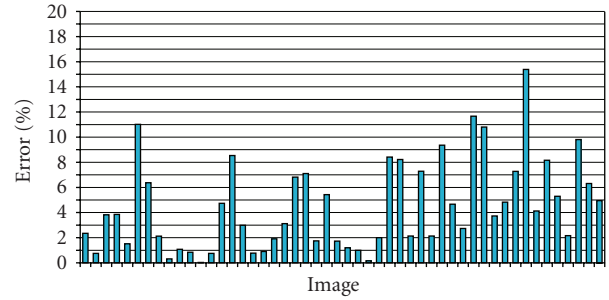
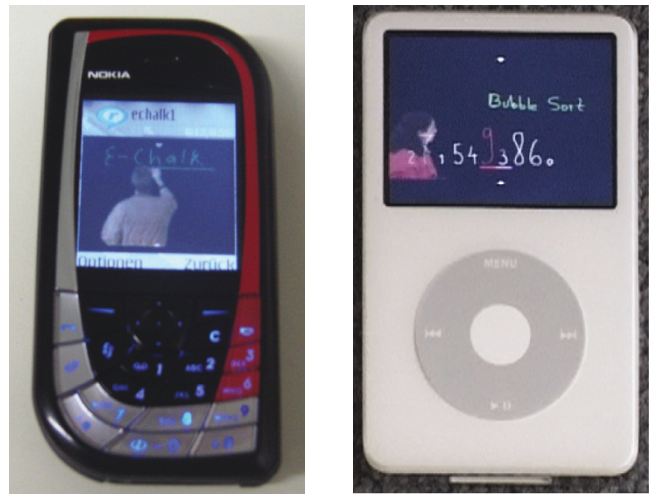


Figure 11: Per-image error measurement from applying SIOX on the benchmark dataset provided by [50]. Please refer to the text for a detailed description.



(a) (b)

Figure 12: Lecture replay using the video capabilities of small devices. (a): a Symbian-OS-based mobile phone. The resolution is 176×144 pixels. (b): a video iPod.

9. RESULTS

The resulting segmented instructor video is scaled to fit the board resolution (usually 1024×768) using linear interpolation. It is pasted over the board content at the receiving end of the transmission or lecture replay. Several examples of lectures that contain an extracted and overlaid instructor can be seen in Figures 4 and 10.

The performance of the presented segmentation algorithm depends on the complexity of the background and on how often it has to be updated. Usually, the current Java-based prototype implementation processes a 640×480 video at 25 frames per second after the initialization phase.

Reflections on the board display are mostly classified as background and small moving objects never make up the biggest connected component. For the background reconstruction process to collect representative background pixels, it is not necessary to record a few seconds without the instructor. The only requirement is that, for the first few seconds of initialization, the lecturer keeps moving and does not occlude background objects that differ significantly from those in the other background regions.

As the algorithm focuses on the background, it provides rotation and scaling invariant tracking of the biggest moving object. The tracking still works when the instructor turns around or when he leaves the scene and a student comes up to work on the board. Once initialized, the instructor does not disappear, even if he or she stands absolutely still for several seconds (which is actually very unusual).

10. FORMAL EVALUATION

A generalized version of the algorithm has been published under the name SIOX (Simple Interactive Object Extraction, http://www.siox.org). It can be used for various segmentation tasks and has been implemented as a low-interaction still-image segmentator into the open source image manipulation programs GIMP and Inkscape. A detailed evaluation of the robustness of the approach including benchmark results can be found in [40, 51].

In order to evaluate the strengths and weaknesses of the color signature segmentation approach more formally, we benchmarked the method using a publicly available benchmark. In [52], a database of 50 images plus the corresponding ground truth to be used for benchmarking foreground extraction approaches is presented. The benchmark data set is available on the Internet [50] and also includes 20 images from the Berkeley Image Segmentation Benchmark Database [53]. The data set contains color images, a pixel-accurate ground truth, and user-specified trimaps. The trimaps define a known foreground region, a known background region, and an unknown region. We chose comparison with this database because the solutions presented in [52] are the basis for the so-called "GrabCut" algorithm, which is commonly considered to be a very successful method for foreground

extraction (though not fast enough for real-time video processing). Unfortunately, this way we cannot test the motion statistics part of our approach (described in Section 7.2) because the benchmark only concerns still images. However, the motion statistics part is relatively simple and straightforward and never turned out to be an accuracy bottleneck.

The error measurement in [52] is defined as

$$\epsilon = \frac{\text{no. misclassified pixels}}{\text{no. of pixels in unclassified region}}.$$

If both background and foreground k-d trees are built, the best-case average error of the algorithm is 3.6%. If only the background signature is given (as presented in this article), the overall error is 11.32 %. The best-case average error rate on the database reported in [52] is 7.9%. The image segmentation task defined in the benchmark exceeds by far the level of difficulty of our segmentation task. Yet, we get reasonable results when using this benchmark.

11. CONCLUSION

This article proposes changing the way chalkboard lecture webcasts are to be transmitted. The standard side-by-side replay of video and blackboard content causes technical and cognitive problems. We propose cutting the lecturer image out of the video stream and pasting it on the rendered representation of the board. The lecturer—a human being—is brought back to the remote lecturing scenario so each remote lecture becomes "human-centered" or "anthropocentric" instead of handwriting-centered. Our experiments show that this approach is feasible and also aesthetically appealing. The superimposed lecturer helps the student to better associate the lecturer's gestures with the board contents. Pasting the instructor on the board also reduces space and resolution requirements. This makes it also possible to replay a chalkboard lecture on mobile devices (see Figure 12).

ACKNOWLEDGMENTS

The E-Chalk system is an ongoing project at Freie Universität Berlin since 2001. Several others have contributed to the system, including Kristian Jantz, Christian Zick, Ernesto Tapia, Mary-Ann Brennan, Margarita Esponda, Wolf-Ulrich Raffel, and—most noticeably—Lars Knipping.

REFERENCES

[1] M. Gleicher and J. Masanz, "Towards virtual videography," in *Proceedings of the 8th ACM International Conference on Multimedia (MULTIMEDIA '00)*, pp. 375–378, ACM Press, Los Angeles, Calif, USA, October-November 2000.

[2] Y. Rui, L. He, A. Gupta, and Q. Liu, "Building an intelligent camera management system," in *Proceedings of the 9th ACM International Conference on Multimedia (MULTIMEDIA '01)*, vol. 9, pp. 2–11, ACM Press, Ottawa, Canada, September-October 2001.

[3] M. Wallick, R. Heck, and M. Gleicher, "Marker and chalkboard regions," in *Proceedings of Computer Vision/Computer Graphics Collaboration Techniques and Applications (Mirage '05)*, pp. 223–228, INRIA Rocquencourt, France, March 2005.

[4] G. D. Abowd, "Classroom 2000: an experiment with the instrumentation of a living educational environment," *IBM Systems Journal*, vol. 38, no. 4, pp. 508–530, 1999.

[5] R. Anderson, R. Anderson, O. Chung, et al., "Classroom presenter—a classroom interaction system for active and collaborative learning," in *Proceedings of the 1st Workshop on the Impact of Pen-based Technology on Education (WIPTE '06)*, West Lafayette, Ind, USA, April 2006.

[6] R. Rojas, G. Friedland, L. Knipping, and E. Tapia, "Teaching with an intelligent electronic chalkboard," in *Proceedings of the ACM SIGMM Workshop on Effective Telepresence (ETP '04)*, pp. 16–23, New York, NY, USA, October 2004.

[7] R. Krauss, R. Dushay, Y. Chen, and F. Rauscher, "The communicative value of conversational hand gestures," *Journal of Experimental Social Psychology*, vol. 31, no. 6, pp. 533–552, 1995.

[8] M. G. Riseborough, "Physiographic gestures as decoding facilitators: three experiments exploring a neglected facet of communication," *Journal of Nonverbal Behavior*, vol. 5, no. 3, pp. 172–183, 1981.

[9] W. Hürst and R. Müller, "The AOF (authoring on the fly) system as an example for efficient and comfortable browsing and access of multimedia data," in *Proceedings of the 9th International Conference on Human-Computer Interaction Education (HCI '01)*, pp. 1257–1261, New Orleans, La, USA, August 2001.

[10] C. Dufour, E. G. Toms, J. Lewis, and R. Baecker, "User strategies for handling information tasks in webcasts," in *Proceedings of the Conference on Human Factors in Computing Systems (CHI '05)*, pp. 1343–1346, ACM Press, Portland, Ore, USA, April 2005.

[11] S. D. Kelly and L. Goldsmith, "Gesture and right hemisphere involvement in evaluating lecture material," *Gesture*, vol. 4, no. 1, pp. 25–42, 2004.

[12] A. Fey, "Hilft Sehen beim Lernen: Vergleich zwischen einer audiovisuellen und auditiven Informationsdarstellung in virtuellen lernumgebungen," *Unterrichtswissenschaften, Zeitschrift für Lernforschung*, vol. 4, pp. 331–338, 2002.

[13] U. Glowalla, "Utility und Usability von E-Learning am Beispiel von Lecture-on-demand Anwendungen," *Fortschritt-Berichte VDI*, vol. 22, no. 16, pp. 603–621, 2004.

[14] J. Sweller, P. Chandler, P. Tierney, and G. Cooper, "Cognitive load as a factor in the structuring of technical material," *Journal of Experimental Psychology*, vol. 119, no. 2, pp. 176–192, 1990.

[15] P. Chandler and J. Sweller, "The split attention effect as a factor in the design of instruction," *British Journal of Education Psychology*, vol. 62, pp. 233–246, 1992.

[16] R. Mertens, G. Friedland, and M. Krüger, "To see or not to see: layout constraints, the split attention problem and their implications for the design of web lecture interfaces," in *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, & Higher Education (E-Learn '06)*, pp. 2937–2943, Honolulu, Hawaii, USA, October 2006.

[17] P. Kellman, *Ontogenesis of Space and Motion Perception*, Academic Press, New York, NY, USA, 1995.

[18] G. Cooper, "Cognitive load theory as an aid for instructional design," *Australian Journal of Educational Technology*, vol. 6, no. 2, pp. 108–113, 1990.

[19] J. C. Tang and S. L. Minneman, "Videodraw: a video interface for collaborative drawing," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '90)*, pp. 313–320, ACM Press, Seattle, Wash, USA, April 1990.

[20] J. C. Tang and S. Minneman, "Videowhiteboard: video shadows to support remote collaboration," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*

(CHI '91), pp. 315–322, ACM Press, New Orleans, La, USA, April-May 1991.

[21] N. Roussel, "Exploring new uses of video with videospace," in *Proceedings of the 8th IFIP International Conference on Engineering for Human-Computer Interaction (EHCI '01)*, pp. 73–90, Springer, Toronto, Canada, May 2001.

[22] M. Apperley, L. McLeod, M. Masoodian, et al., "Use of video shadow for small group interaction awareness on a large interactive display surface," in *Proceedings of the 4th Australasian User Interface Conference (AUIC '03)*, pp. 81–90, Australian Computer Society, Adelaide, Australia, February 2003.

[23] A. Tang, C. Neustaedter, and S. Greenberg, "Videoarms: embodiments for mixed presence groupware," in *Proceedings of the 20th British HCI Group Annual Conference (HCI '06)*, London, UK, September 2006.

[24] A. Tang, C. Neustaedter, and S. Greenberg, "Embodiments for mixed presence groupware," Tech. Rep. 2004-769-34, Department of Computer Science, University of Calgary, Calgary, Canada, 2004.

[25] S. Gibbs, C. Arapis, C. Breiteneder, V. Lalioti, S. Mostafawy, and J. Speier, "Virtual studios: an overview," *IEEE Multimedia*, vol. 5, no. 1, pp. 18–35, 1998.

[26] R. Gonzalez and R. Woods, *Digital Image Processing*, Prentice Hall, Upper Saddle River, NJ, USA, 2nd edition, 2002.

[27] L. Li and M. K. H. Leung, "Integrating intensity and texture differences for robust change detection," *IEEE Transactions on Image Processing*, vol. 11, no. 2, pp. 105–112, 2002.

[28] A. Elgammal, D. Harwood, and L. Davis, "Non-parametric model for background substraction," in *Proceedings of the 7th IEEE International Conference on Computer Vision, Frame Rate Workshop (ICCV '99)*, Kerkyra, Greece, September 1999.

[29] N. Friedmann and S. Russel, "Image segmentation in video sequences: a probablistic approach," in *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence (UAI '97)*, Providence, RI, USA, August 1997.

[30] M. Simon, S. Behnke, and R. Rojas, "Robust real time color tracking," in *RoboCup 2000: Robot Soccer World Cup IV*, pp. 239–248, Springer, Melbourne, Australia, August-September 2001.

[31] I. Haritaoglu, D. Harwood, and L. S. Davis, "W^4: real-time surveillance of people and their activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 809–830, 2000.

[32] D. Beymer, P. McLauchlan, B. Coifman, and J. Malik, "A real-time computer vision system for measuring traffic parameters," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '97)*, pp. 495–501, San Juan, Puerto Rico, USA, June 1997.

[33] L. Li, W. Huang, I. Y. H. Gu, and Q. Tian, "Foreground object detection from videos containing complex background," in *Proceedings of the 11 ACM International Conference on Multimedia (MULTIMEDIA '03)*, pp. 2–10, Berkeley, Calif, USA, November 2003.

[34] S. Jiang, Q. Ye, W. Gao, and T. Huang, "A new method to segment playfield and its applications in match analysis in sports video," in *Proceedings of the 12th ACM International Conference on Multimedia (MULTIMEDIA '04)*, pp. 292–295, ACM Press, New York, NY, USA, October 2004.

[35] S.-Y. Chien, Y.-W. Huang, S.-Y. Ma, and L.-G. Chen, "Automatic video segmentation for MPEG-4 using predictive watershed," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME '01)*, pp. 941–944, Tokyo, Japan, August 2001.

[36] J. Y. A. Wang and E. H. Adelson, "Representing moving hands with layers," *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 625–638, 1994.

[37] Y. Wang, T. Tan, and K.-F. Loe, "Video segmentation based on graphical models," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '03)*, vol. 2, pp. 335–342, Madison, Wis, USA, June 2003.

[38] M. A. Nascimento and V. Chitkara, "Color-based image retrieval using binary signatures," in *Proceedings of the ACM Symposium on Applied Computing (SAC '02)*, pp. 687–692, ACM Press, Madrid, Spain, March 2002.

[39] B. C. Ooi, K.-L. Tan, T. S. Chua, and W. Hsu, "Fast image retrieval using color-spatial information," *The VLDB Journal*, vol. 7, no. 2, pp. 115–128, 1998.

[40] G. Friedland, K. Jantz, and R. Rojas, "SIOX: simple interactive object extraction in still images," in *Proceedings of the 7th IEEE International Symposium on Multimedia (ISM '05)*, pp. 253–259, Irvine, Calif, USA, December 2005.

[41] G. Friedland, K. Jantz, T. Lenz, F. Wiesel, and R. Rojas, "A practical approach to boundary-accurate multi-object extraction from still images and videos," in *Proceedings of the 8th IEEE International Symposium on Multimedia (ISM '06)*, pp. 307–316, San Diego, Calif, USA, December 2006.

[42] CIE, "Recommendations on Uniform Color Spaces, Color-Difference Equations, Psychometric Color Terms," Supplement No. 2 of CIE Publication No. 15 (E-1.3.1) 1971, 1978.

[43] G. Wyszecki and W. S. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulae*, John Wiley & Sons, New York, NY, USA, 1982.

[44] E. Hering, *Outlines of a Theory of the Light Sense*, Harvard University Press, Cambridge, Mass, USA, 1964.

[45] L. Hurvich and D. Jameson, "An opponent-process theory of color vision," *Psychological Reviews*, vol. 64, pp. 384–404, 1957.

[46] CIE, "Colorimetry (Official Reccomendations of the International Commision on Illumination)," CIE Publication No. 15 (E-1.3.1), 1971.

[47] B. Hill, Th. Roger, and F. W. Vorragen, "Comparative analysis of the quantization of color spaces on the basis of the CIELAB color-difference formula," *ACM Transactions on Graphics*, vol. 16, no. 2, pp. 109–154, 1997.

[48] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[49] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[50] Microsoft Research, "Microsoft Foreground Extraction Benchmark Dataset," 2004, http://research.microsoft.com/vision/cambridge/i3l/segmentation/GrabCut.htm.

[51] G. Friedland, *Adaptive audio and video processing for electronic chalkboard lectures*, Ph.D. thesis, Department of Computer Science, Freie Universität Berlin, Berlin, Germany, 2006.

[52] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr, "Interactive image segmentation using an adaptive GMMRF model," in *Proceedings of the 8th European Conference on Computer Vision (ECCV '04)*, vol. 3021 of *Lecture Notes in Computer Science*, pp. 428–441, Springer, Prague, Czech Republic, May 2004.

[53] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings of the 8th IEEE International Conference on Computer Vision (ICCV '01)*, vol. 2, pp. 416–423, Vancouver, BC, Canada, July 2001.