Abstract

Practical video scene and face recognition systems are sometimes confronted with low-resolution (LR) images. The faces may be very small even if the video is clear, thus it is difficult to directly measure the similarity between the faces and the high-resolution (HR) training samples. Face recognition based on traditional super-resolution (SR) methods usually have limited performance because the target of SR may not be consistent with that of classification, and time-consuming SR algorithms are not suitable for real-time applications. A new feature extraction method called coupled kernel embedding (CKE) is proposed for LR face recognition without any SR preprocessing. In this method, the final kernel matrix is constructed by concatenating two individual kernel matrices in the diagonal direction, and the (semi)positively definite properties are preserved for optimization. CKE addresses the problem of comparing multimodal data that are difficult for conventional methods in practice due to the lack of an efficient similarity measure. Particularly, different kernel types (e.g., linear, Gaussian, polynomial) can be integrated into a uniform optimization objective, which cannot be achieved by simple linear methods. CKE solves this problem by minimizing the dissimilarities captured by their kernel Gram matrices in the LR and HR spaces. In the implementation, the nonlinear objective function is minimized by generalized eigenvalue decomposition. Experiments on benchmark and real databases show that our CKE method indeed improves the recognition performance.

Keywords: Image Recognition, Low-Resolution, Hybrid Digital and Coupled Kernel Algorithm.

INTRODUCTION

Human motion tracking is primarily concerned with determining the existence and location of humans within certain regions of space. Automatic face recognition is a process of identifying a test face image with one of the faces stored in a prepared face database. Real world images need not necessarily contain isolated face(s) that can directly serve as inputs to a FR system. Hence, there is a need to isolate or segment facial regions to be fed to a FR system. Most of the time, a video sequence of the scene is available using which a person may have to be recognized.

For recognition, we need the face position in which it is best recognizable by the present day FR algorithms. Hence, a robust system that detects and tracks a face is necessary. Face detection and tracking becomes an important task with the growing demand for content-based image functionality. Though human beings detect/track faces with very little effort, it is not easy to train a computer to do so. In pattern recognition parlance, human face is a complex pattern.

Different poses and gestures of the face accentuate complexity. The detection scheme must operate flexibly and reliably regardless of the lighting conditions, background clutter in the image, multiple faces in the image, as well as variations in face scale, pose and expression. The system should be able to detect the face even under small occlusions. Over the past several decades, surveillance techniques have matured dramatically. Analog taps and security personnel are being replaced with Internet Protocol (IP) technology, leveraging digital video cameras, remote access, and intelligent analytics. This evolution provides organizations with significant opportunities to improve security and reduce operating costs.

Today’s businesses and public agencies are faced with a critical need to protect employees, clients, citizens and assets from possible threats with a security system that enables rapid response to security breaches and prompt investigation of events. Organizations are additionally challenged with managing tremendous amounts of information in various forms, including video, voice, electronic data and paper. The goal of this technological and commercial intelligence report is to describe the intelligent video surveillance sector for the security of individuals and places. It is an emerging, little known technology that is changing how traditional video surveillance is used and is opening up a world of opportunities, making it possible to foresee new market segments emerging in the security sector. This document, intended for a non-expert audience, discusses the ins and outs of this technology and tries to characterize the market it represents, not only globally, but more specifically in Quebec. It contains information on video surveillance technology, its application, shift to IP networks, leading-edge video analytic techniques applicable to it, its needs, the developments and trends in this field, the issues it raises, and the supply and demand it generates.

Video surveillance is a segment of the physical security industry, which also includes access control, fire detection and control, the technical management of buildings, systems to ensure individual safety and the detection of intrusion. Video surveillance consists of remotely monitoring public or private places, using mostly power-operated cameras that transmit the images taken to monitoring equipment that records or reproduces the images on a screen. It captures images of moving people in order to monitor comings and goings, prevent theft, assault and fraud, as well as manage incidents and crowd movements.
Video analytics, also called intelligent video surveillance, is a technology that uses software to automatically identify specific objects, behaviours or attitudes in video footage. It transforms the video into data to be transmitted or archived so that the video surveillance system can act accordingly. It may involve activating a mobile camera in order to obtain more specific data about the scene or simply to send a warning to surveillance personnel so that a decision may be made on the proper intervention required. Intelligent video surveillance systems use mathematical algorithms to detect moving objects in an image and filter non-relevant movements. They create a database that records the attributes of all the objects detected and their movement properties. Decisions are made by the system or events of interest are searched in archived footage based on rules (e.g., if a person oversteps a boundary, send an alert).

The video captured by surveillance cameras must be sent to the recording, processing and viewing systems. This transmission can be done by cable (coaxial or fibre optic cables, stranded copper wire) or by air (infrared signals, radio transmission). Wired video is the most predominant in video surveillance systems. It offers greater bandwidth and better reliability than wireless connections, at a lower cost. However, wireless video transmission is sometimes the best solution, for example when monitoring large perimeters where installing cables would be too costly, or when the areas to be monitored cannot be reached by cable.

Whether wired or wireless, the video signal can be analogue or digital. Most video transmissions for surveillance are currently still analogue. However, computer networks (LAN, WAN or Internet) are increasingly used to send video using the IP protocol. IP cameras can be directly connected on these networks, whereas the video flow coming from analogue cameras must first be digitized by an encoder, also called a video server, in order to pass through the IP networks. Video management systems process video surveillance images, such as managing different video flows, and viewing, recording, analyzing and searching recorded footage. There are four major categories of video management systems.

**Digital Video Recorder (DVR22):**

Device with an internal hard disk for digital recording of video and built-in video processing software. It accepts only flows from analogue cameras, which it digitizes. Recent models make it possible to view the video remotely on computer. Still quite widespread, it is slowly being replaced by systems that support IP video from end to end.

- **Hybrid Digital Video Recorder (HDVR23):**

  Similar to the digital recorder, but accepts connections from both analogue and IP cameras. Several types of digital video recorders can be made hybrid by installing a software application.

- **Network Video Recorder (NVR24):**

  Designed for video surveillance IP network architectures, it can only process video signals from IP cameras or encoders.

- **IP video surveillance software:**

  Purely software-based solution for managing video on an IP network. For surveillance systems with few cameras, a Web browser may be enough to manage the video. For larger video surveillance networks, a dedicated video management software application must be used, which is installed on a PC or server. Although more complicated to install due to the required server configurations, it offers greater flexibility with respect to choice and the addition of video surveillance network parts. IP video surveillance software applications are a major trend in video management, especially in infrastructures with large numbers of cameras. Open platforms allow for easy integration of cameras and hardware components from different manufacturers.

A VIDEO camera produces images at a certain frame rate and depth of field, which impose physical limits on the spatial density of image detectors. Except for expensive professional-quality video cameras, video cameras can produce only low-resolution (LR) videos. Practical face recognition systems are confronted with LR problems, and the performance of the system usually declines when the input face images are degraded [1]–[3], such as low resolution of only $\delta \times 8$ pixels. Compared with high-resolution (HR) images, these LR images lose some discriminative details across different subjects.

**OBJECTIVES**

1. To prevent theft from happening both indoors and outdoors of the premises. The storage facilities and delivery areas within multiplexes are easy targets of small time thieves who are lurking for some fast cash. With the right supervision you can monitor your staff and keep in control that only authorized personnel enter the office.
2. While watching a movie or playing a video game there are chances of it getting ugly over a feud. These security cameras are a deterrent to all the violent outburst and fights that might occur.
3. Every business is running in the world due to the strong customer base that it has, if they only are the prime target then the direct impact is to the business. Ensuring that your customers are your prime important entity and their security is well monitored is important.
4. Damage to your belongings can be done by a person you trust to catch him right in action a security camera is your ideal tool. It ensures work ethics among your staff.
5. We cannot undermine the importance of this footage that is captured on tape as this is quite valuable in court of law. This video of any culprit committing a crime is a proof enough to put him behind the bars.
6. A new feature extraction method called coupled kernel embedding (CKE) is proposed for LR face recognition without any SR preprocessing.
7. In this method, the final kernel matrix is constructed by concatenating two individual kernel matrices in the diagonal direction, and the (semi)positively definite properties are preserved for optimization.

**REVIEW OF PREVIOUS WORK**

Recent research on algorithms have attempted to avoid explicit SR in the image domain. The approach performs SR reconstruction in the eigen face domain has been investigated in Gunturk et al.(2003) work. The researchers propose a video-based LR face recognition approach with implicit SR (Arandjelovic and R. Cipolla ,2007) More recently(P. Hennings-Yeomans et al., 2008) proposed a joint objective function that integrates the aims of super-resolution and face recognition. Compared with two-step methods, this method improves the recognition rate. However, the speed of this algorithm is slow, even for the speed-up version, because the optimization procedure needs to be executed for each test image with regard to each enrollment.

Most of the existing work that addresses the problem of
matching faces across changes in pose and illumination cannot be applied when the gallery and probe images are of different resolutions. The commonly used approach for matching a low resolution (LR) probe image with a high resolution (HR) gallery is to use super-resolution to construct a higher resolution image from the probe image and then perform matching. There have been a few recent efforts that address recognition and super-resolution simultaneously (Hennings-Yeomans, et.al,2008).

Blanz and Vetter (2003) propose a 3D morphable model based approach in which a face is represented using a linear combination of basis exemplars. The shape and albedo parameters of the model are computed by fitting the morphable model to the input image. Romdhani et al.(2002) provide an efficient and robust algorithm for fitting a 3D morphable model using shape and texture error functions. Zhang and Samaras(2006) combine spherical harmonics illumination representation with 3D morphable models (2003). An iterative approach is used to compute albedo and illumination coefficients using the estimated shape. For face recognition across pose, local patches are considered more robust than the whole face, and several patch based approaches have been proposed; (Jia and Gong,2005). In a recent paper, Prince et al.(2008) propose a generative model for generating the observation space from the identity space using an affine mapping and pose information.

Baker and Kanade (2003) proposed an algorithm to learn a prior on the spatial distribution of the image gradients for frontal facial images. Chakrabarti et al. (2007) proposed a learning-based method using kernel principal component analysis for deriving prior knowledge about the face class for performing super-resolution. Liu et al. (2007) proposed a two-step statistical modeling approach for hallucinating a HR face image from a LR input. The relationship between the HR images and their corresponding LR images is learned using a global linear model and the residual high-frequency content is modeled by a patch-based non-parametric Markov network. Chang et al. (2004) use manifold learning approaches for recovering the HR image from a single LR input. Yang et al. (2010) address the problem of generating a super-resolution image from a LR input image from the perspective of compressed sensing. A novel patch-based face hallucination framework is proposed by Liu et al. (2005).

METHODOLOGY

**Existing System**

- A video camera produces images at a certain frame rate and depth of field, which impose physical limits on the spatial density of image detectors.
- Normally the video captured by the camera are of low resolution.
- So, the images in the video may be small and of low resolution images.
- So, it is very much difficult to directly measure the similarity between the images and High Resolution images.
- Except for expensive professional-quality video cameras, video cameras can produce only low-resolution (LR) videos.
- Practical face recognition systems are confronted with LR problems, and the performance of the system usually declines.

- SR may not be consistent with that of classification, and time-consuming SR algorithms are not suitable for real-time applications.
- So, Face recognition based on traditional super-resolution (SR) methods usually have limited performance.

**Proposed System**

- A new feature extraction method called coupled kernel embedding (CKE) is proposed for LR face recognition without any SR preprocessing.
- The final kernel matrix is constructed by concatenating two individual kernel matrices in the diagonal direction, and the (semi) positively definite properties are preserved for optimization.
- CKE addresses the problem of comparing multimodal data that are difficult for conventional methods in practice due to the lack of an efficient similarity measure.
- Different kernel types (e.g., linear, Gaussian, polynomial) can be integrated into a uniform optimization objective, which cannot be achieved by simple linear methods.

A multi linear approach to hallucinate faces across multiple modalities (generalization to variations such as facial expression or pose) based on a unified global and a local tensor space representation. A joint objective function that integrates the aims of SR and face recognition, which indeed improves the recognition rate.

Discover the nonlinear relationship between the LR subspace and HR subspace with a kernel regression (KR) approach. The KR method consists of two phases, namely, training and recognition. In the training phase, given a set of LR and HR image pairs, both images are mapped to the kernel feature space by nonlinear mappings.

A kernel-subspace-based regression model is designed to estimate the relationship for SR from recognition perspective. In the recognition phase, given the query image, its HR image representation in kernel space can be generated using the learned relationship. Then, any kernel-based face recognition methods, such as Kernel direct discriminative analysis, can be employed for recognition.

**Software Configuration**

**Hardware Requirements**

<table>
<thead>
<tr>
<th>SYSTEM</th>
<th>Pentium IV 2.4 GHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>HARD DISK</td>
<td>40 GB</td>
</tr>
<tr>
<td>FLOPPY DRIVE</td>
<td>1.44 MB</td>
</tr>
<tr>
<td>MONITOR</td>
<td>15 VGA colour</td>
</tr>
<tr>
<td>MOUSE</td>
<td>Logitech.</td>
</tr>
<tr>
<td>RAM</td>
<td>256 MB</td>
</tr>
<tr>
<td>KEYBOARD</td>
<td>110 keys enhanced.</td>
</tr>
</tbody>
</table>

**Software Requirements**

- Operating system: Windows XP Professional
- Front End: Microsoft Visual Studio .Net 2005
- Coding Language: C#.Net
- Database: SQL Server
The block diagram gives a high-level overview of the software architecture of the human motion detector. Input to the system is provided by a video source which can be either a IEEE 1394 digital camera or a sequence of still image files in JPEG format. The IEEE 1394 digital camera input allows for live video to be processed in real-time as it is captured by the camera.

The still image input allows pre-recorded data to be processed. This pre-recorded data can be from any source, such as a Mini-DV camcorder or a video capture card, as long as it is first converted to JPEG format. Video data from the video input is made available to the background subtraction algorithm, which is responsible for differentiating between foreground and background image regions. The implementation of the background subtractor is discussed in more detail. The output of the background subtractor for each video frame is an 8-bit per pixel bitmap that serves as a foreground mask: Those pixels which are foreground have value 255 and those pixels which are background have value 0. The foreground mask provided by the background subtractor is processed to build a data structure that enumerates the boundaries for each distinct foreground object. This enumeration is achieved by the contour finder, a simple algorithm that finds each “connected component” of the foreground mask.

A connected component is defined as a region of an image whose pixels all have the same value and are adjacent to other pixels in the same connected component. For our contour finder, we used an implementation of the algorithm provided by the Intel Open Computer Vision Library. Once these contours are located, those with small geometric area are ignored. Such small contours are likely to be noise or small image disturbances that are probably not human.

The feature tracker outputs a list of point coordinates within this bounding box and a velocity for each. This list of point features is then input to the statistical model detector. The output of the model detector is a single number, the evaluation of the model’s probability density function.

Larger probabilities indicate that the set of input points match the model well, and smaller probabilities indicate a poor match. This probability can be thresholded to decide whether or not the object is human. Finally, once each object is evaluated to be a human or not, the results are rendered on-screen, overlaid on top of the input video. A box is drawn around those objects which are detected to be human. Other visual indicators are also drawn to indicate the status of each stage of the algorithm.

ANALYSIS

Frame Recording

A digital video recorder (DVR), sometimes referred to by the merchandising term personal video recorder (PVR), is a consumer electronics device or application software that records video in a digital format to a disk drive, USB flash drive, SD memory card or other local or networked mass storage device. The term includes set-top boxes (STB) with direct to disk recording facility, portable media players (PMP) with recording, recorders (PMR) as camcorders that record onto Secure Digital memory cards and software for personal computers which enables video capture and playback to and from a hard disk.

Video encoders, also called video servers, are used to convert signals from analogue cameras and transmit them in IP flow to a network through a switch. While retaining the analogue cameras, they enable a practically complete shift to the network infrastructure for video surveillance since the video is constantly transmitted by IP.
protocol through the network.

Video encoders may be used with network video recorders (NVR). These can only process and record IP video flow. They are offered on an open platform (a computer with video management software) or in dedicated proprietary equipment. In this latter form, the network video recorder is compared to a hybrid digital recorder that requires encoders to operate with analogue cameras. This module mainly deals with recording the videos. Using these videos the unauthorized person is tracked. The videos are captured using the webcam.

**Frames Conversion**

There are times in a production when the footage only exists in one format and the user needs to use it in another. A common example of that is a 24p project being handed an establishing shot that was shot on digibeta. In this case, the footage is interlace, 60 fields at 30 frames per second while the project type is progressive and 24 frames per second. There are a number of approximations and simplifications that have been made to increase speed. Most importantly, the timing of the algorithm has been changed so that a new background model is only generated once every 240 frames, or 8 seconds, rather than re-computing the model for each frame. During the time between model generation, statistics for the background are collected from each frame and processed incrementally. Once 240 frames have been reached, a new background model is computed for use on the next 240 frames. Furthermore, since successive frames tend to be very similar, only every fourth frame is used for statistics collection. Another simplification which increases speed is to only perform background subtraction on a sub-sampled version of each image. This module deals with converting the video to frames. The video is converted to frames based on time seconds.

**Coupled Kernel Algorithm**

Kernel-based learning machines have aroused considerable interest in the fields of pattern recognition and machine learning. CKE algorithm implemented on a binary classification problem. There are two modes including HR objects (squares and five-pointed stars) and their LR counterparts (triangles and circles) in the input space, where the squares and triangles belong to the first class, and the five-pointed stars and circles belong to the second class.

**Complexity Analysis**

The computation of CKE for features extraction involves three steps: kernel matrices generation, weights computation, and generalized eigenvalue decomposition. We use the term flam, which is a compound operation consisting of one addition and one multiplication, to measure the operation counts.

**Face Recognition**

The detection and tracking of faces and facial features in video sequences is a fundamental and challenging problem in computer vision. This research area has many applications in face identification systems, model-based coding, gaze detection, human-computer interaction, teleconferencing, etc. This module deals with recognizing the face of the persons in the recorded video and indicate whether they are authorized or unauthorized persons. The tracking formula, is used to compute the displacement for each iteration of the algorithm.

A multi linear approach to hallucinate face images across multiple modalities (generalization to variations such as facial expression or pose) based on a unified global and a local tensor space representation. A joint objective function that integrates the aims of SR and face recognition, which indeed improves the recognition rate. Discover the nonlinear relationship between the LR subspace and HR subspace with a kernel regression (KR) approach. The KR method consists of two phases, namely, training and recognition. In the training phase, given a set of LR and HR image pairs, both images are mapped to the kernel feature space by nonlinear mappings. A kernel-subspace-based regression model is designed to estimate the relationship for SR from recognition perspective. In the recognition phase, given the query image, its HR image representation in kernel space can be generated using the learned relationship. Then, any kernel-based face recognition methods, such as Kernel direct discriminative analysis, can be employed for recognition.
Conclusion

An algorithm has been developed to detect and track human face(s) in a color image sequence. The algorithm starts with human skin color modeling and uses it in isolating skin pixels (probable face regions). Skin color is found to be a powerful feature for isolating potential face candidates. It is also useful for detecting multiple human faces in an image. It is orientation Independent. Connected Component Operators are applied on the threshold skin probability image to isolate the final face components. The combination of the six operators used proved to be very effective. Skin color analysis followed by the use of shape based Connected Operators makes the system invariant to change in scale.

For a higher detection performance, the structuring element used during open/close operations must be changed adaptively. For tracking, a simple but effective approach has been followed. This involves projecting the face regions in the present frame as markers to the next frame, and detecting for skin in the localized regions. This reduces a lot of computational overhead by avoiding face detection in every frame. And since the min-max box dimensions to be projected into the next frame are increased adaptively based on the face size in the present frame, the tracking step is also scale independent. Building a more robust skin model using larger number of skin and non-skin pixels would enhance the performance of the detector.

We have proposed a new approach for LR face recognition without any SR preprocessing. Our method is based on CKE, where the objective function aims to preserve the locality between neighborhood in the reproducible kernel Hilbert space. CKE addresses the problem of comparing multimodal data which is difficult for conventional methods in practice due to the lack of efficient similarity computation. CKE solves this problem by minimizing the inconsistency between the similarity measures captured respectively by their kernel Gram matrices in the two spaces. The experiments using nonlinear kernel functions on real surveillance camera databases show the effective improvement in recognition accuracy.

Future Enhancement

Future work includes pose estimation and 3D head tracking by using the positions of pupils, nostrils and lip corners. Furthermore, we intend to improve the system robustness developing a recovery module and considering motion to increase the accuracy of detection and tracking of facial features. High-level video understanding can be performed based on images taken from a single static camera and with simple perception methods working almost in real-time. This has been possible by using two sets of a priori information: first, contextual information describing the 3D geometry of the observed scene and semantic information on the static objects and interesting areas, second, general knowledge of predefined scenarios valid for an application domain. The current video understanding framework we propose has shown several limitations. One type of problems is the imprecision and uncertainty in the detection and location of mobile objects; most of these low-level detection errors are due either to reflections, shadows or occlusions.

A solution to cope with these problems is to relax our second hypothesis and not to restrict ourselves to the use of a single camera. Another more general problem is that as every vision system, this framework needs, for each perception method and for each interpretation method, to set the values of numerical parameters in a configuration phase. One solution to solve this problem is to use learning techniques to find the best parameter values for an application.

There are a variety of enhancements that could be made to this system to achieve greater detection accuracy and increased robustness:

- Objects could be tracked between frames rather than simply performing human motion detection on single frames. For example, a Kalman filter could be used to predict the future position and human likelihood of a given object. Such a filter would smooth out particular frames in which detection fails, and would eliminate many false detections. The net effect would be an improvement in both PD and PFA along with a smooth tracking capability useful for higher-level applications.
- As described above, the current model of motion does not take into account the time dependent nature of a walking human. Much greater accuracy would be possible with a detector and model that takes advantage of this periodicity in time.
- The current background subtraction algorithm can be confused by fast lighting changes or moving shadows. A better algorithm would use a technique based on optical flow for the image segmentation. This approach would also allow the camera to be in motion relative to the background.
- Modeling different types of human motion should be explored, such as walking seen from different viewpoints. The current system fails to detect humans walking directly towards or away from the camera. In addition, other forms of motion such as running should be modeled so that the detector can reliably detect and classify these cases.

References

on Computer Vision and Pattern Recognition, 478–484.

Theory and practice. International Journal of Computer Vision,

analysis for face recognition across large pose differences. IEEE
Trans. on PAMI, 30(6):970–984.

fitting a 3d morphable model using linear shape and texture error

13) Yang, J, J. Wright, T. Huang, and Y. Ma (2010). Image super
resolution via sparse representation. IEEE Trans. on Image
Processing.

training image under arbitrary unknown lighting using spherical