Video Face Tracking and Recognition with Skin Region Extraction and Deformable Template Matching

Simon Clippingdale, NHK Science & Technology Research Laboratories, Japan
Mahito Fujii, NHK Science & Technology Research Laboratories, Japan

ABSTRACT

The authors describe a face tracking and recognition system for video indexing that handles variable face poses (left-right and up-down) and deformations due to speech and facial expressions. The system is based on deformable template matching, and employs person-specific templates at near-frontal poses for recognition, and novel person-independent templates at multiple poses on the view-sphere for tracking. Relative to an earlier version that used multiple person-specific templates at multiple (left-right) poses, the new system speeds up processing by (i) restricting attention to skin-color regions; (ii) performing recognition using the person-specific templates at near-frontal poses only; and (iii) tracking at non-frontal poses using the novel person-independent templates. Registration is also simplified since multiple views of each target individual are no longer required, but at the cost of a loss of recognition functionality at poses far from frontal (the system instead “remembers” the identity of each individual from near-frontal matches and tracks between them).

Keywords: Deformable Template Matching, Face Recognition, Face Tracking, Multimedia Indexing, Skin Color Detection, Video Archive Indexing

1. INTRODUCTION

The tracking and recognition of faces in video offers the prospect of annotating video archives with face metadata that could subsequently be used for automatic searching and retrieval. Estimation of facial expressions, deformations associated with speech movements, and speaker and speech recognition could potentially extend the scope of this indexing to semantically richer metadata that would allow the rapid subsequent retrieval of scenes on the basis of human-meaningful descriptions such as “the scene where person A says X to person B,” or “all scenes where person A talks about subject X,” or “the sequence where person A smiles at the interviewer.”

A number of technical challenges must be overcome for any such video or multimedia indexing system to be practical. A system that
is intended to have general applicability must handle the many degrees of freedom found in typical video, including moving cameras and backgrounds, 3D head rotation, lighting, dynamic occlusions, and facial deformations associated with speech movements and facial expressions. Extraction and recognition of these deformations could contribute even richer semantics (Yeasin et al., 2006).

Less flexible systems can still give reasonable results on suitably constrained video. The NameIt system (Satoh et al., 1999) attempts to identify near-frontal faces in news video. It first detects faces using the neural network based face detector of Rowley et al. (1998), and then matches them to candidate person names extracted either from closed captions or from the audio track by speech recognition. Face similarity across different shots (and across sequences where the face appears at non-frontal pose) is assessed by an eigenface based method (Turk & Pentland, 1991).

The Character Appearance Retrieval and Analysis (CARA) system (Jung et al., 2006) developed by Korean Broadcasting System (KBS) extracts near-frontal faces with a Viola-Jones cascade-style face detector (Viola & Jones, 2004) and attempts to match them based on discrete cosine transform features. The system has been deployed in a production setting.

Systems designed to assist commercial and home photo and video editing have begun to use face information. Home users are now familiar with face tagging on social media sites, and are prepared to correct mistaken suggestions produced by face recognition software. Apple has released a product incorporating face processing, although its performance may suffer on unconstrained scenes (Shankland, 2010). Employing multiple cues including soundtrack to improve reliability has been a common approach, adopted by systems including the fully-automatic video editing application Magisto (http://www.magisto.com/). Another system designed for home use (Takiguchi, Adachi, & Ariki, 2010) relies on speech detection to spot scenes of interest, and then attempts to perform facial expression recognition.

However, few systems have attempted to deal with the variability in face images found in unconstrained video as described. Systems that the authors have developed are intended to be a step in that direction. We originally developed a flexible, deformable-template-based face tracking and recognition system (Clippingdale & Ito, 1999; Clippingdale & Fujii, 2003) that identifies a set of facial feature points in the target video. It matches the feature point positions (i.e., shape or “where”) and Gabor wavelet features measured at those positions (i.e., texture or “what”) against corresponding feature point positions and Gabor wavelet features contained in a number of deformable templates that are constructed from images of target individuals at the registration stage. That “original” system attempted to handle head rotation by registering templates of each target individual at multiple head poses on a left-right axis. As the face region in the input video rotated left or right away from frontal pose, the best match would be obtained by one of the non-frontal registered templates. On each input face region, matches were attempted with templates at poses within an interval around the pose estimated in the previous frame.

A number of issues limited the practical applicability of the original system: (i) it did not handle up-down head rotation; (ii) matching against templates for all registered individuals (within a range of poses) was computation-intensive; (iii) it was often difficult to obtain imagery for registration showing target individuals at the required set of left-right poses (and adding up-down head rotations would have compounded this problem even further). Additionally, the system overall was too computationally intensive to attain reasonable (near real time) speed with standard PC hardware. Some extensions of the original system to statistical matching methods (Matsui et al., 2004, 2006) and 3D registration (Clippingdale et al., 2009) offered improved performance in various respects, but at the cost of even more intensive computation.

With a view to practical applications in the short-to-medium term, a new version of
Multimodal Information Integration and Fusion for Histology Image Classification
www.igi-global.com/article/multimodal-information-integration-fusion-histology/54462?camid=4v1a