Vitals are Vital! Visual Vital Metric Detection for Multiple Subjects

Enhao Gong
Electrical Engineering
Stanford University
enhaog@stanford.edu

Jason Bouhenguel
School of Medicine
Stanford University
jjbouh@gmail.com

Abstract

Vitals are vital! Patient vitals provide critical information essential to the successful diagnosis and management of patients in every healthcare settings, particularly in the acute-care setting. Heart-rate(HR), and to a greater degree heart rate variability, as well as overall patient mobility can indicate the presence of severe disease states. Practical and technical constraints currently limit monitoring in healthcare settings to select (eg. admitted) patients, leaving numerous others unmonitored and potentially at risk. Wed like to use Computer vision to provide a non-invasive solution to this acute problem.

To tackle this problem, we extended Haar cascade detectors and SIFT based tracking and face recognition for a robust continuous face detection and assignment. Applying independent component analysis (ICA) based blind signal decomposition we were able to accurately extract heart-rate from video. Additionally, we proposed a method based on multi-scale spatial-temporal correlation in motion field to recognize illness related activities (eg. coughing).

The HR detection error is $1 \pm 2bpm$ *and action recognition (coughing) is* 90%*. This work could be used to triage patients in an unmonitored, possibly acute, healthcare setting. Further improvement of the algorithms and implementation is in progress.*

1. Introduction

1.1. Motivation

A common proverb pervasive throughout medical training is vitals are vital. The utility of patient vitalsspecifically heart-rate [1], blood pressure, respiratory rate, temperaturein the acute healthcare setting is unrivaled by much. Simple objective parameters like a patients' heart-rate can tell us when a patient is experiencing extreme pain or anxiety [2]; it can alert us to the potential presence of shocky states concerning for internal hemorrhage, severe dehydration, or infections which require immediate attention; it can warn us of the presence of arrhythmia and heart disease; or suggest possible toxin exposure/drug overdose.

While hospital protocol is to have patients on constant monitoring upon admission to the hospital, areas such as ED or clinic waiting rooms, cafeterias, etc are void of any such monitoring. While having constant monitoring of patients in acute medical settings of any kind would be ideal, the reality is that it is not always possible due to a lack of resources (personnel, equipment, etc) or is simply impractical to have patients walking around with monitors affixed to their person. A common mentality in medicine is the less invasive the better; now imagine a world in which monitoring is about as invasive as standing in air (while having your picture/video taken that is).

1.2. Challenges

First of all, a robust face/subject detection and recognition is still unsolved for wild environment involving multiple subjects. We designed and developed a multi-stage framework to tackle this problem.

One current challenge is that available algorithms are sensitive to motion or other severe changes in environment. We tackled these problems and improved the algorithms using computer vision and statistical methods.

An additional challenge is to extend current single- or few-subject algorithms to being able to handle multi-subject detection, recognition, and analysis.

In addition, the algorithm efficiency, stability to motion/resolution and the capability to be real-time were of major concern.

1.3. Goals

Our goal was to make use of "visual vitals" technology, applying it to many subjects simultaneously in a dynamic setting (eg. an emergency department waiting room) with the ultimate goal of identifying high-risk patients that may require immediate attention. Further, combining visually

extractable information with past medical history through integration of patient-specific electronic medical records (as available) would provide a way to contextualize extracted vital information and better assess health status and thereby provide continuous risk-profiling and automated triaging of patients.

2. Related work and comparison

2.1. Review of previous work

Previous work from MIT: Poh et al. (2010) [3, 4] debuted an ICA based algorithm for extracting heart-rate from standard quality video recordings, which was subsequently followed by Wu et al (2012) [5] who described an alternaive approach using a spatio-temporal amplification method for determining an individuals heart-rate from mere video surveillance of the subject. These algorithms provide tools to analyze biometrics from videos using sequential or hierarchical approaches.

2.2. Contribution and advantage of proposed method

2.2.1 Main contributions

The main contributions in our work is:

- designed and developed a multi-subject face detection tracking method robust to motion and blurring.
- developed a blind-signal decomposition based method to robustly extract heart-rate.
- proposed a motion field spatial-temporal correlation based method to extract features for mobility metrics and illness activity recognition.
- proposed and developed a basic system for computer-vision based noninvasive and passive bio-metric monitoring for multiple subjects.

2.2.2 Comparison with the color/motion amplification method

We implemented the algorithm [5] proposed by MITs CSAIL group (Wu et al) in which the spatial-temporal information was magnified. Generally, the algorithm proposes to amplify the changes using a Eulerian (pixel changes) instead of Lagrangian specification (motion tracking) of flow-field. The basic steps include:

- Transform RGB to NTSC (lamination Y and chrominance I and Q)
- Decompose the video YIQ-t 4D data into spatial-temporal pyramid with multiples scales of resolution as a hierarchical structure.
- Filter out the information of interest using temporal filter to the frames
- Magnification of the information using weighted summation
- Compose video again by combining the spatial-temporal pyramid

Implementing the algorithm revealed several susceptibilities, specifically, significant sensitivity to motion and ambient light changes. Also, their use of filter parameters will introduce bias for heart-rate monitoring which makes the method non-ideal for clinical heart-rate detection.

So our main advantages are **robust to subject motion** and **unbiased in frequency selection**.

The same MIT group has recently been proposed another method, using motion information instead of color changes to quantify heart-rate which could be better suited for this application, a choice which slightly coincides with our method.

2.2.3 Comparison with ICA based method

Our approach is similar to the aforementioned ICA based algorithm [3, 4]. We extended it with:

- robust face detection and tracking
- more sophisticated method to localize the peaks in spectrogram
- multiple subject tracking and assign the signal into different profiles
- extended the biometric with mobility metrics and action recognition and potentially with heart-rate Variability.

2.2.4 Comparison with activity recognition method

There has been significant research in the area of fine-grained activity recognition. Our propose method incorporates elements from methods using correlation between spatial and temporal information [6], whereby motion information is explicitly estimated using optical flow instead of implicitly use the location information. This approach is better suited to the specific application of detecting illness related activities, as such activities are more likely to involve quick small-scale movements for which pixel-level resolution of motion estimation of relative positions within a tracking bounding box of body would be ideal.

Also, our multi-subject tracking result can be potentially useful for the recognition of collective activity [7].

3. Technical Part

3.0.5 Technical Part: Summary of technical solution

The main system consists of several modules including:

- 1) robust face detection and tracking using cascade detector, motion field estimator and SIFT
- 2) Heart-rate extraction using ICA based blind-signal decomposition and dynamic spectrogram analysis.
- 3) Mobility metric estimation using optical flow.
- 4) Illness related action recognition using motion field spatial-temporal correlation based features and spatial pyramid dimension reduction.

The basic step for the technical approach is a framework to constantly detect and track multiple subject. We extended the classic Haar cascade detectors with pixel tracking information. Basically, the face detection and new bounding boxes are approximated with efficient tracking results, checked with color histogram metrics, and periodically corrected with SIFT based tracking. The detector constantly detects faces and we recognize the subject using SIFT feature based method and update the bounding boxes. This framework outperforms the classic cascade detectors and facilitates multi-subject tracking, better handling missed detection and motion blurring.

For visual based heart-rate monitoring, we aggregated the RGB signal in the detected bounding boxes and use independent component analysis (ICA) based blind signal decomposition to extract a heart-rate sensitive channel. Based on this channel signal, we use Fourier Transform robustly estimate the heart-rate.

For illness-related action recognition, we generated an optical flow-based motion responses in different directions and at multiple spatial and temporal scales. Through dimensional reduction–using blocks and spatial pyramid–, we generated the *spatial-temporal correlation in motion field* features representing detailed motion information, and then quantitatively estimated the accurate mobility level and recognized illness specific activity using SVM.

3.1. Technical Part: Detail of technical solutions

3.1.1 Robust Face detection and tracking using motion information

The basic method we used for face detection is the classic Haar Cascade Detectors. We used several cascade detectors for frontal faces, profile faces as well as upper bodies to locate subjects as shown figure 1

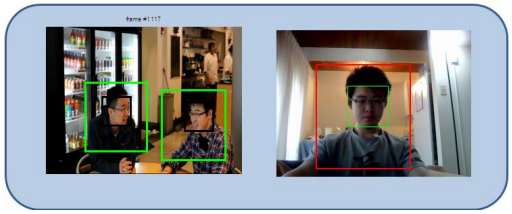

Figure 1. Basic face and body detection

Robust face detection involving multiple subjects in various orientations is still an open problem in computer vision. Figure 2, below, presents a basic flowchart depicting how our system works.

We designed and developed a multi-stage approach to improve the face detection and tracking in which motion field information was used to robustly track faces even in the event of large changes or blurring due to motion. SIFT and histogram based method periodically double-checked that the tracking bounding-box is valid. In the event that a new face is detected by our Cascade detectors, the SIFT based method will recognize the owner of the face, and appropriately update the corresponding bounding box and metrics in the subject's profile. We made several strides to improve the robustness of subject/face tracking. The algorithm is shown in 1.

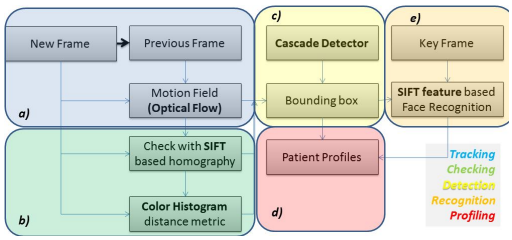

Figure 2. We designed and developed multi-stage approach to improve the face detection and tracking in which: (a) motion field information was used to robustly track faces even if there are large changes or blurring due to motions. (b) SIFT and histogram based method periodically double-checks tha the tracking bounding-box is valid. (c) In the event that a new face is detected using Cascade detectors (e) the SIFT based method will recognize who the face belongs to and will update the corresponding bounding box and the metrics in the (d) subject's profile

Algorithm 1 Proposed steps to face detection and multi-subject tracking

Initialization

a) Detect faces in keyframe as initialization

b) Approximate face bounding-box based on the motion tracking

Tracking:

c) Check the tracked bounding-box is similar to previous frame's to the extant of color histogram, matched SIFT features, etc.

d) Periodically detect faces using Haar Cascade detectors (for frontal face, profile face, bodies)

e) Recognize who the face belongs to, update the corrected bounding box information in their profile (based on tracking and detection results).

f) Interpolate any missing detections

3.1.2 Face Recognize combining SIFT features and eigenfaces

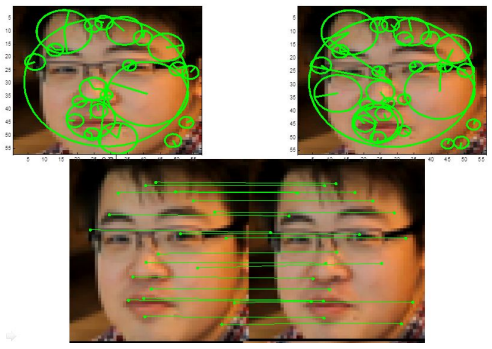

Figure 3. The SIFT keypoints and descriptors are used for face tracking correction and face recognition.

Algorithm 2 Steps of the proposed face recognition

Training

a) Use one (or a few) frames to mimic the check-in image which is a key frame with easily localizable faces

b) Detect faces and define bounding boxes

c) Compute SIFT keypoints and features in the detected bounding boxes. Save into the profile for future face recognition

d) When a new face is detected, compute SIFT keypoints and features

e) Find the matches between SIFT features from keyframe and new frame.

Continuous Learning

f) Based on SIFT results, align images and improve the eigen-face PCA based methods.

Output:

g) Recognized face/patients' IDs and update the SIFT features/eigenface features.

Here we used Scale Invariant Feature Transform (SIFT) descriptors in two step.

In the first step, we periodically use SIFT descriptor to compare the image in the face bounding box in two consecutive frames. The keypoints ($n_{keypoint} = 500$) were detected and matched based on SIFT descriptor using RANSAC.

If there are too few valid matches the detected bounding box will be replaced; this happens when a subject moves out of the scene and motion field cannot locate the face anymore.

In the second step, we use SIFT information for initial face recognition and bounding box assignment. We compute robust face feature vectors using SIFT features from the key frames as initial features and save them into a subject profile database. When a new face is detected, the SIFT features in the new face are compared with those in candidate faces. Histogram distance (via the fast Quadratic Chi metric) was used as the criteria. Once more faces are detected, we can use the SIFT to align them together and use Linear Discriminant Analysis (LDA) based method for more robust recognition.

The main algorithm is shown in 2.

3.1.3 Robust heart-rate detection

Our visual heart-rate detection algorithm was designed in the image of that presented by Poh et al (2010).

- 1) For each facial ROI identified by our detection algorithm, frame-by-frame spatial averages of respective encircled pixels were calculated in a channel-wise (Red, Green, Blue) manner.
- 2) These face-specific RGB(t) average values were sub-selected using a 30-second sliding window [(30 seconds)x(Frame Rate) number of frames], and normalized across each color channel, respectively.
- 3) ICA was performed (using FastICA library) followed by FFT (2048 samples) on all three resulting components.
- 4) Power spectra were analyzed and the most prominent harmonic (corresponding to an individuals heart beat) within our physiologic window (0.75 to 4Hz::45 to 240bpm) was selected. Peak selection was further constrained to include only those frequencies within 0.2Hz of the frequency corresponding to the prior measured heart-rate (1 second prior), if available. This takes advantage of physiologic hear-rate changes being typically less than 12bpm over a one second time interval.
- 5) The 30-second window was slid 1 second [(1second)x(Frame Rate) number of frames] and steps 2-5 were repeated until the signal was completely analyzed.

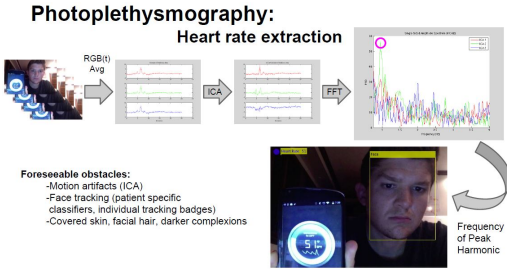

Figure 4. Method for heart-rate monitoring using blind signal decomposition and spectral analysis

3.1.4 Mobility Metrics

Patient movement behavior can be useful for diagnosis of certain disease states and as an overall indicator of patient well-being. Here we used an optical flow based method to extract the motion field of motion vectors for each pixel in different frames. Based on the estimate motion field, we can quantitatively estimate how much each subject has moved in the video. This turns out to be a useful metric to quantify mobility and activeness of subjects.

3.1.5 Illness-related activity detection

In addition to whether the subjects are moving, we can actually learn more about the activities of subject using the accurate motion field information.

Here we designed and developed a method extracting motion features and using them to train a SVM classifier to recognize a person coughing.

The features extraction part is similar to the *Spatial-Temporal Correlation* but instead of the correlation in key-word domain we exploit the spatial-temporal correlation in the motion field domain. The dynamic motion field provide accurate pixel level of motions. Here we filtered the motion field with different spatial and temporal filters (smoothing). Then we can get motion responses in multiple spatial and temporal scales. We proposed to use these responses as features since they are good representations for fine-grained categories of actions.

For example, when the subject coughs there will be little motion in large temporal scale with rapid bulk (large spatial scale)

motions. For speaking, instead, one observes small spatial scale motion in small temporal scales.

We used a spatial pyramid method to reduce the dimension of features and train the SVM classifer with RBF kernels based on the manual tag of whether there are cough in each frames. 5-fold cross-validation was used to evaluate the performance.

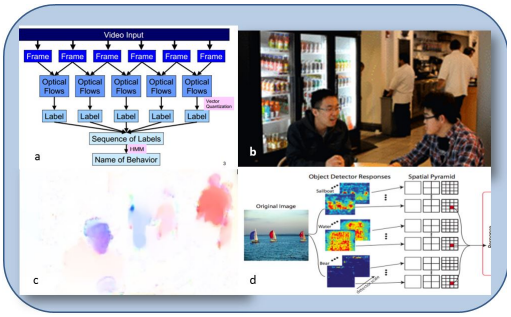

Figure 5. Method for motion feature extraction (1) using optical flow and filtering with multiple spatial-temporal resolution, which is a form of the spatial-temporal correlation in motion field, from (2) frames with illness related motion into (3) motion field responses. (4) Spatial pyramid based method can further reduce the dimension for the feature used in SVM

4. Experiments

4.1. Experiment design and data materials

4.1.1 Video collection

We collected several video datasets with single or multiple subject for test the model and improve the algorithm. We took the video from a 720x480 web-cam from a laptop with frame rate of about 17fps and also a 1088x1920 HD Camera with tripod to take videos with frame rate of about 30fps. The dataset include subject with little motions and moving a lot in certain in-room environment.

All the dataset used is around 1 minute to 5 minute based on the need of signal processing. The video processing was conducted in Matlab which was offline instead of real-time online implementation.

4.1.2 Experements for illness related activity

Here we designed algorithm to recogize illness related activities.

Using cough as an example of proof-of-concept, videos with subject's occasionally deliberately coughing was collected to estimate the performance.

4.1.3 Comparison with ground-truth

We used standard plethysmography to accurately measure heart-rate during video capture. The plethysmograph was self-constructed using an Arduino Mico micro controller, paired with a light intensity-to-voltage converter (TAOS TSL267 High-Sensitivity IR Light-to-Voltage Converter), and red emitting LED (680 nm).

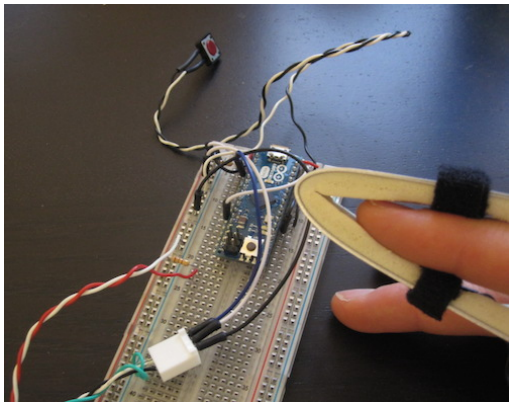

Figure 6. The plethysmograph was self-constructed using an Arduino Mico micro controller, paired with a light intensity-to-voltage converter (TAOS TSL267 High-Sensitivity IR Light-to-Voltage Converter), and infrared emitting LED (940 nm).

Raw data (figure 7) was read over serial port and post-processed in Matlab using standard FFT and peak selection techniques, akin to those used subsequently for heart-rate extraction from video.

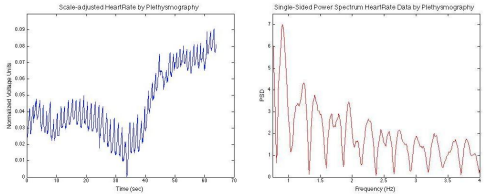

Figure 7. Plethysmograph raw data was read over serial port and post-processed in Matlab using standard FFT and peak selection techniques.

4.2. Results

4.2.1 Summary of results

Using the above methods, we implemented Computer Vision algorithm to constantly and passively detect vital metrics for multiple subject in the scene. The HR detection error is $1 \pm 2bpm$ and the action recognize (for cough) is 90%.

The metric can be used to triage patients with specific risk model to help healthcare providers. The methods can be applied in other scenarios as well, such as restaurants, classrooms and retails. The performance can be further improved, and the development of more integrated software and more robust algorithm is still in progress.

4.2.2 Multi-subject Tracking

Here we implement the method with multiple video dataset in which there are multiple subject moving and sitting. The algorithm can detect and track the subjects in almost all the frames.

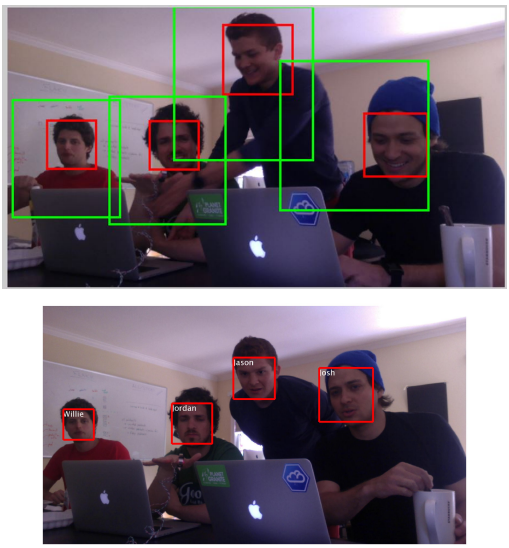

Figure 8. Robust detection, tracking (above) and recognition (below) for multiple subjects in wild scenario.

4.2.3 Heart-rate estimation

Based on the comparison with estimated HR from video and aligned sensor signal, we can evaluated the systematic and random error of the method.

We analyzed the Fourier Transform for the sensor signal and selected the principal peak as the goundtruth heart-rate.

The result from a video with single person, comparing the groundtruth with our proposed computer vision based estimation from video, the error was:

$$error_{HR} = 1.1 \pm 2.1bpm$$

Figure 9 shows the comparison between groundtruth and the video estimation from a video shown in 8. The error is worse than the result from a video with single person and more pixels for the face. We can see that the trends are largely matched and the differences are bounded.

4.2.4 Mobility Metrics and Illness related activity Recognition

The estimation of motion is related to the parameter for motion field estimation. Currently, we are using multi-resolution method for optical flow which can achieve pixel level accuracy. So the mobility metric is with **1cm-5cm** level accuracy. Here we used 5-fold cross-validation for testing the recognition of illness related activity. The overall accuracy for recognition is 90.0%.

5. Discussion

5.1. Performance of the algorithm

5.1.1 Improved detection with blurred image

Since we approximate face detection with optical flow based tracking first, we can estimate the motion field and how face/subject

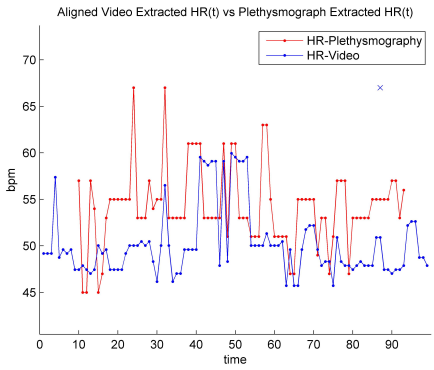

Figure 9. Evaluation of the performance of HR monitoring from video. The estimated HR from video and from sensor signal were alignment based on timestamp and the Root-Mean-Square-Error was computed. The systematic error was $4.4bpm$ while the random error was $3.9bpm$.

moves with pixel-level accuracy. Thus, we can get a good updated face bounding box even in the presence of blurring artifacts– a source of failure for many contemporary detection algorithms. An example is shown in Figure 10.

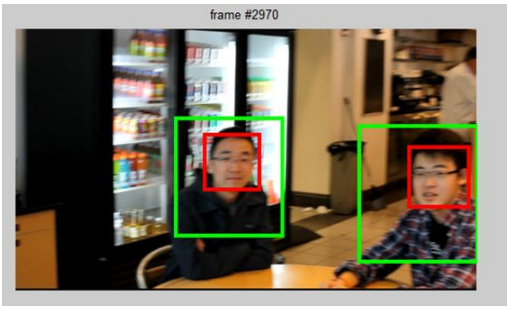

Figure 10. The figure demonstrate a correct detection in highly blurred images. The proposed multi-stage algorithm can more robustly detect and track faces and subjects even if the images are blurred due to the subject's movement.

5.2. Further improvement

5.2.1 Improved HR signal using clustering

Here we estimated the HR related signal with the aggregated signal in RGB channels within the entire bounding box. Instead, we can cluster the pixel in bounding box first to segment out which of them are from face instead of background or hairs. Preliminary results show this would improve the accuracy in HR monitoring. Incorporating temporal Wavelet transforms into our heart-rate detection algorithm may also provide an avenue to increase overall performance.

5.3. Risk management model and methods

In this project, we extracted heart-rate and mobility vital metrics from videos.

Variations in heart-rate are proportional to the averaged heart-rate in sliding windows, which is inversely proportional to the averaged R-R interval (Heart Beat Interval). Mathematically, the variance of the estimated Heart Beat Interval is related to the Heart Rate Variability (HRV) and the duration of the selected sliding window. HRV is a clinically vital metric which provides an estimate of cardiac autonomic modulation. In recent years the utility of HRV as a predictor of numerous disease states has expanded–specifcally correlating with mortality after myocardial infarction, congestive heart failure, diabetes, depression, poor survival in premature babies, and brain death [8]. Specific to our interests here, many have propposed its use as a means to assess and triage patients in the Emergency Department.

The mobility represent the activity of the patients and is also highly related to illness and mortality [9]. The illness related activity, cough for example, is a more direct indicator of diseases.

Additionally, incorporating other imaging modalities such as infrared could provide us with access to additional vitals such as temperature and regional blood perfusion.

While we were unable to implement EMR integration currently, contextualizing patient vitals with medical history through will provide us with an opportunity to better risk stratify and triage patients in a busy patient-heavy setting.

5.4. Applications

The proposed method and system can also be applied in various scenarios. For example, retail sales and restaurants can use this to further analyze the video data they have as an input to find the nervousness, anxiety and emotion of each customers to further boost their business.

Figure 11. Applications

6. Conclusion

Using the above methods, we implemented CV algorithm to noninvasively detect vital metrics for multiple subject in the scene. The HR detection error is $1 \pm 2bpm$ and the action recognize (for cough) is 90%.

The metric can be used to triage patients with specific risk model to help healthcare providers. The methods can be applied in other scenarios as well, such as restaurants, classrooms and retails. The performance can be further improved, and the development of more integrated software and more robust algorithm is still in progress.

6.1. References

References

[1] Juul Achten and Asker E Jeukendrup. Heart rate monitoring. *Sports medicine*, 33(7):517–538, 2003.

[2] Bruce H Friedman and Julian F Thayer. Autonomic balance revisited: panic anxiety and heart rate variability. *Journal of psychosomatic research*, 44(1):133–151, 1998.

[3] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express*, 18(10):10762–10774, 2010.

[4] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *Biomedical Engineering, IEEE Transactions on*, 58(1):7–11, 2011.

[5] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics (TOG)*, 31(4):65, 2012.

[6] Silvio Savarese, Andrey DelPozo, Juan Carlos Niebles, and Li Fei-Fei. Spatial-temporal correlatons for unsupervised action classification. In *Motion and video Computing, 2008. WMVC 2008. IEEE Workshop on*, pages 1–8. IEEE, 2008.

[7] Wongun Choi, Khuram Shahid, and Silvio Savarese. Learning context for collective activity recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3273–3280. IEEE, 2011.

[8] Mark L Ryan, Chad M Thorson, Christian A Otero, Thai Vu, and Kenneth G Proctor. Clinical applications of heart rate variability in the triage and assessment of traumatically injured patients. *Anesthesiology research and practice*, 2011, 2011.

[9] Mirja Hirvensalo, Taina Rantanen, and Eino Heikkinen. Mobility difficulties and physical activity as predictors of mortality and loss of independence in the community-living older population. *Journal of the American Geriatrics Society*, 48(5):493–498, 2000.