SOMIB Sociedad Mexicana de Ingeniería Biomédica

43° CNIB 2020

GESTO-MX: Construcción a Distancia de un Dataset de Expresiones en Rostros Mexicanos

J. F. Navarro[1]*
[1]Universidad Autónoma de Baja California, Baja California, México
*navarro.juan@uabc.edu.mx

Resumen

Este trabajo muestra la construcción de un conjunto de herramientas para la captura a distancia, compilación, preprocesamiento y clasificación de una base de datos de imágenes de expresiones en rostros mexicanos. Se realizó la captura de 5 distintas expresiones faciales de 75 individuos a través de una aplicación web, produciendo un conjunto de datos de alrededor de 9000 imágenes. A estas imágenes se les aplicó una ecualización de histograma, transformación afín y recorte de región de interés. Se realizó una comparación de distintos algoritmos de aprendizaje máquina para la clasificación de los gestos, obteniendo una precisión máxima de 95.8 % para el gesto de alegría y 75.0 % para el gesto de enojo. El conjunto de datos desarrollado a través de este estudio, está disponible bajo petición, bajo el nombre de GESTO-MX. Las herramientas para su construcción quedaron disponibles en el dominio público, con el fin de que investigadores de toda la comunidad científica mexicana pueda aportar a la construcción de una base de datos de expresiones de rostros con fenotipos característicos de la población mexicana. Por otro lado está la posibilidad de desarrollar otros experimentos a distancia con estas herramientas, para elaborar conjuntos de datos o desarrollar algoritmos de aprendizaje máquina.

Palabras Clave: Aprendizaje máquina, Captura a distancia, Dataset, Emociones, Gestos

1. Introducción

Los conjuntos de datos (más comúnmente llamados *dataset*) de imágenes de rostros y expresiones faciales son utilizados para entrenar algoritmos de detección y predicción que permiten la implementación de todo tipo de aplicaciones prácticas. Aunque existen bastantes conjuntos de imágenes de gestos faciales [1] [2], esta área no debería estar exenta de continua exploración, ya que todos presentan inconvenientes que dificultan ciertas aplicaciones, por ejemplo: en el caso de los datasets JAFFE, Yale, CK, CK+, MMI y KDEF, existen menos de 600 imágenes en cada uno [1], al ser tan limitada la cantidad de muestras, los modelos generados con estas presentarían dificultad para generalizar. Por otro lado, aunque los datasets AR, MUG y TFEID son muy completos en sus condiciones experimentales y cantidad de muestras (en el rango de los miles), presentan el inconveniente de que los dos primeros contienen casi exclusivamente rostros de ascendencia europea y el último asiática [1], siendo estos también obstáculos para la generalización o la aplicación específica en el campo mexicano.

Recientemente se ha hecho investigación para mitigar este desbalance en la representatividad fenotípica en conjuntos de gestos faciales, como es el caso del dataset RADIATE [3], desafortunadament no se han publicado esfuerzos similares que busquen la *representatividad de rostros latinoamericanos, y mucho menos mexicanos*, en la investigación de visión computacional.

La necesidad de abordar las limitaciones que trae esta deficiencia se vuelve evidente considerando que existen varias aplicaciones prácticas para el reconocimiento de expresiones faciales en áreas de la salud, tales como la detección de expresiones de dolor [4], sistemas auxiliares para el control de dispositivos en individuos con discapacidades físicas [5], sistemas de atención para conductores [6], medición del efecto de traumas psicológicos, además de su uso en conjunto con otros sistemas de captura

dx.doi.org/10.24254/CNIB.20.11

de señales fisiológicas [7], como medidores de conductancia de la piel [8], electrocardiograma [9], razón respiratoria [10], entre muchos otros, aumentando su precisión y poder predictivo [11]. Debido a que al implementar alguna de esta aplicaciones es necesario entrenar los sistemas con un dataset grande de rostros, y la gran mayoría de imágenes estructuradas disponibles presentan *en minoría o en proporción casi nula rostros mexicanos y/o latinoamericanos*, este podría ser una limitante que no permita llegar a una meta de máxima precisión en el despliegue de algunas de estas aplicaciones en nuestro país.

El objetivo del presente estudio es construir a distancia la primera versión de un conjunto de datos de expresiones faciales de rostros mexicanos, que presente una gran cantidad de imágenes, además de publicar herramientas robustas que permitan a otros investigadores cooperar en su mejora continua.

2. Metodología

2.1. Participantes y Sistema de Captura

La población de participantes se conformó por 75 individuos mexicanos de varias edades, siendo en su mayoría estudiantes universitarios (entre 18 y 22 años de edad), que participaron voluntariamente en el estudio anunciado por medio de redes sociales. Todos fueron informados sobre el experimento y qué se haría con sus datos. Se les dió un conjunto de instrucciones con el fin de maximizar la calidad de las imágenes, junto con un aviso de privacidad que aceptaron. En una aplicación web de nombre `gesto-app`, desarrollada para este estudio, se capturaron imágenes del rostro de los participantes realizando 5 expresiones faciales distintas. Los participantes entraron al sitio en internet desde su propio computador con cámara web y la aplicación en línea funcionó como entorno experimental que utilizó dicha cámara como sistema de captura.

Fig. 1. Disposición del experimento, se aprecia la interfaz de la aplicación web.

2.2. Descripción del Experimento

Los participantes, sentados cómodamente frente a la pantalla de su computador, recibieron instrucciónes de cómo realizar el experimento, empezando con la activación de su cámara web. La interfaz en pantalla está dividida en dos mitades, presentando así dos secciones laterales. En la mitad izquierda se aprecia el rostro del participante capturado en vivo (figura 1) y en la mitad derecha aparecen en texto instrucciones básicas con el fin de decirle al participante qué hacer a continuación y un botón para avanzar en cuanto esté listo para realizar la acción indicada. Se hicieron 5 ensayos con cada participante, donde en cada uno se le dió la instrucción realizar una expresión determinada: una expresión de prueba (libre), un rostro neutral, una expresión de alegría, una de tristeza y una de enojo, etiquetadas *test*, *neutral*, *happy*, *sad*, y *angry* respectivamente en el dataset. Se decidió capturar 4 emociones con el fin de que la duración del experimento no fuera demasiado larga, ya que los participantes al ser voluntarios, podrían

no disponer con mucho tiempo y paciencia. Estas 4 emociones son universales y las más comunes en casi todos los datasets de expresiones faciales [1]. En el caso de la expresión *test*, aunque fue incluida en el conjunto de datos compilado, su propósito es servir como un ensayo de prueba para que el participante se adapte al formato del experimento, y no cometa errores en subsiguientes ensayos. Se le solicitó al participante mover su rostro suavemente en varias direcciones con el fin de capturar una variedad de orientaciones faciales. En cada ensayo se realizó la captura automática de 25 imágenes. Algunos pocos participantes cerraron la aplicación a la mitad o en algún otro momento del experimento, quedando registrados únicamente los ensayos que logró terminar. El experimento tiene una duración aproximada de 4 minutos.

2.3. Preprocesamiento de los Datos

Se desarrollo un código `gesto-scraper`, que toma los datos de salida de `gesto-app` y realiza el preprocesamiento y compilado del dataset. Se realizaron las siguientes operaciones de preprocesamiento sobre las imágenes capturadas.

2.3.1. Ecualización del Histograma Adaptativa

Esta técnica es usada frecuentemente para el balanceo de condiciones de iluminación en todo tipo de imágenes. En el estudio se transformó la imagen (figura 2.1) al espacio de color (Y, C_R, C_B) (figura 2.2). Tomando únicamente la componente Y de intensidad, se graficó la frecuencia de sus observaciónes en un histograma (visible en la 2.3) y posteriormente se redistribuyeron los valores a través del histograma logrando que este fuera más uniforme, y menos localizado (figura 2.4), mejorando el contraste del canal Y de intensidad (figura 2.5). Dicho canal mejorado fue agrupado con los canales C_R y C_B no modificados y se realizó la conversión de vuelta al espacio RGB (figura 2.6), dando origen a una imagen de mayor contraste (visible en la figura 2.7). Dicha ecualización ocurrió al nivel de una cuadrícula de subgrupos de 8x8 pixeles, es decir, se aplicó la distribución del histograma de cada subgrupo, para que existiese un mayor contraste local, a este proceso se le llama ecualización del histograma *adaptativa*.

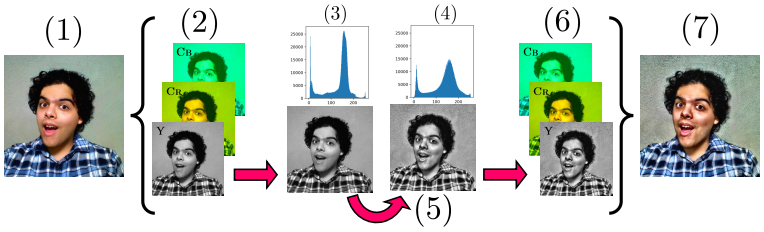

Fig. 2. Ecualización del histograma adaptativa, como ocurrió en el preprocesamiento.

2.3.2. Alineación y Recorte de las Imágenes

Se utilizó un predictor de rostros frontales de la librería `OpenCV` [12], para preservar sólo las imágenes donde se detectó un rostro (figura 3.1). Utilizando un modelo previamente entrenado de la libreria `dlib`, se detectaron 68 puntos de referencia (descritos más adelante en la figura 5.1) en las imágenes de los rostros, de los cuales se utilizaron los puntos de referencia 9, 37 y 46 (visibles en la figura 3.2) para realizar una transformación afín que relocalizó las tres coordenadas a lugares predeterminados en una nueva imagen (visible en la figura 3.3). Los puntos de referencia que denotan el contorno del rostro (figura 3.4), se utilizaron para trazar un polígono de región de interés (figura 3.5) y utilizarlo como

máscara para recortar la imágen (figura 3.6), dando como resultado una imagen de 250×250 pixeles, que incluye solamente el recorte del rostro de la persona, con sus ojos y mentón alineados, sobre un fondo negro (visible en la figura 3.7).

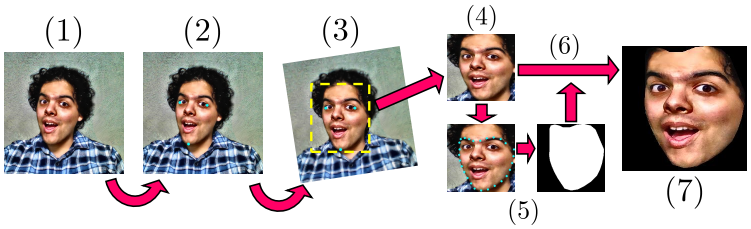

Fig. 3. Recorte de las imágenes, como ocurrió en el preprocesamiento.

2.4. Compilación del Dataset

Después de aplicar las operaciones anteriores al conjunto de imágenes, estas se agruparon en un árbol de directorios, por participante y por categoría, como se observa en la figura 4. Los datos personales fueron removidos del conjunto y los directorios de los participantes fueron etiquetados como `subject1` hasta `subject75`. Dicho ordenamiento fue aleatorio y no es indicativo del orden de los experimentos.

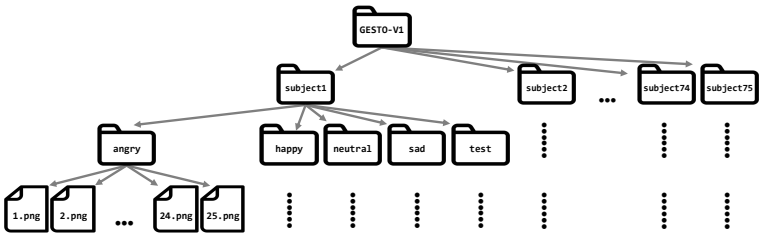

Fig. 4. Árbol de directorios del dataset, en la forma que quedó ordenado.

2.5. Extracción y Clasificación

Se tomaron las imágenes del conjunto en todas las categorías, a excepción de la categoría *test*, previamente preprocesadas y se aplicaron distintas técnicas de extracción de características y algoritmos de aprendizaje máquina para su clasificación, con el fin de obtener una línea base de rendimiento con algunos métodos comunes.

2.6. Técnicas de Extracción de Características

Análisis de Componentes Principales (PCA) aplicado al conjunto de todas las imágenes, se tomaron los primeros 200 componentes (ya que presentaron 95 % de la varianza) de cada una. Este método se utilizó, ya que es comúnmente aplicado para la extracción de características de rostros [13] y expresiones faciales [14].

Landmark Distance (LMD) con base en el método propuesto por F. Z. Salmann, et al. [15], este consistió en medir la distancia pitagórica entre varios puntos del rostro (figura 5.1 y 5.2). Visibles en la figura 5.3, se hicieron los cálculos necesarios para obtener 11 componentes, estos fueron: la distancia entre el vértice interno de los ojos y el extremo interno de sus respectivas cejas ($D1$ y $D7$), la medida de la apertura de ambos ojos ($D2$ y $D8$), la diferencia de altura de cada ojo con sus extremos de la boca ($D3$ y $D9$), la distancia entre la punta de la nariz y el centro del labio superior ($D4$), distancia entre los vértices externos de los ojos ($D11$), además de la apertura y el ancho de la boca ($D5$ y $D6$). A diferencia del trabajo de F. Z. Salmann, et al [15], se utilizaron 5 distancias más, del mismo tipo, aunque reflejadas hacia la otra mitad del rostro, con el fin de abarcarlo completamente.

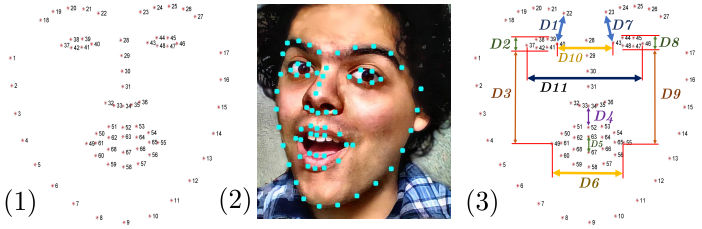

Fig. 5. Puntos de referencia como son entregados por el modelo de `dlib` (1), contextualizados en un rostro (2) y las 11 distancias ($D1$ hasta $D11$) utilizadas (3).

Método Propuesto (LMD+PCA) Por último, se propuso una combinación de los dos métodos anteriores, que consiste en la concatenación de los 11 componentes de LMD y de los primeros 60 componentes de PCA, como características extraídas de cada imagen. La razón principal es debido a que el método de PCA resultó bueno para la detección del rostro *happy* y el método de LMD para el de *angry*, por lo que se consideró que una combinación podría aprovechar la información provista por ambos métodos.

2.7. Técnicas de Clasificación

Se utilizaron 5 técnicas distintas de clasificación, con la características en común de ser de bajo costo computacional y carecer de hiperparámetros. Para la medición de su desempeño, fue necesario particionar el conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba, los cuales tomaron las proporciones de 80 % y 20 % del conjunto respectivamente. La separación ocurrió de forma estratificada con respecto a los participantes, no al total de las imágenes, esto con el fin de que la medición de las métricas se diese de forma válida, ya que imágenes distintas de la misma categoría de un sólo participante son muy similares, pudiendo causar sobreajuste de los modelos.

Máquina de Vectores de Soporte de tipo Lineal y de tipo Funcion de Base Radial (SVM Lineal y RBF) utilizando el enfoque `one-vs-rest`, ya que con respecto a la detección de rostros, este es un enfoque muy efectivo [16], además de que SVM es el método más utilizado en este tipo de aplicaciones [1].

Árboles de decisión (DT) optimizados bajo el criterio de GINI, estos han demostrado en la literatura ser buenos para la clasificación. En el caso del estudio [15], se logró un rendimiento de 92 % de precisión para el rostro de alegría y 70 % en el de enojo utilizando el dataset CK+, lo cual es muy importante tomando en cuenta la baja complejidad matemática y computacional de este tipo de modelos.

Análisis de Discriminante Lineal (LDA) y Cuadrático (QDA). El análisis de discriminante lineal, también conocido como análisis discriminante de Fisher, es similar con el método utilizado de SVM, pues separa linealmente las categorías, auqnue en lugar de consistir en una optimización con respecto a vectores de soporte, obtiene una solución analítica al asumir una condición de normalidad en la distribución de los datos. Si esta condición de normalidad es verdadera y existen pocas muestras atípicas, LDA puede ser óptimo. QDA es muy similar, pero maneja un contorno de decisión más complejo, pudiendo tal vez ajustarse mejor a algunos conjuntos de datos.

3. Resultados

3.1. Recursos

A través de esto se desarrollaron tres recursos sencillos de utilizar, que se encuentran a la disposición de cualquier investigador que tenga la intención de realizar aportaciones a este esfuerzo (en la dirección `https://github.com/biogylo/gesto-tools/`).

- Una aplicación web, `gesto-app`, desarrollada en `JavaScript`. Una vez cargada en un servidor, solicita datos de contacto de los participantes que ingresen al sitio y tras la aceptación de un aviso de privacidad conforme a la ley, utiliza la webcam para capturar imágenes del participante realizando ciertas expresiones faciales (un ejemplo de la interfaz se observa en la figura 1), con el fin de realizar la compilación de un dataset.
- La primera versión de un dataset, GESTO-MX, compilado el 25 de Julio del 2020, formado con la participación de 75 participantes mexicanos en su mayoría estudiantes universitarios, que presentaron 4 ensayos donde realizaron una expresión facial diferente en cada uno en distintas orientaciones y 1 ensayo de prueba, capturando 25 imágenes diferentes en cada uno, produciendo en conjunto, alrededor de 9 000 imágenes. La información para descargar dicho dataset estará disponible en: `https://github.com/biogylo/gesto-tools/tree/master/gesto-dataset` y en caso de no estar arriba el sitio, puede ser solicitada al correo electrónico del autor.
- Un código, `gesto-scraper`, escrito en `Python`. Este toma las imágenes capturadas por la aplicación web y aplica una serie de pasos de preprocesamiento a cada una, para después sortearlas en un árbol de directorios por participantes, y tipo de expresión facial, con el fin de compilar el dataset.

3.2. Desempeño de los Métodos de Clasificación

Utilizando el algoritmo QDA y el método de extracción propuesto de PCA+LMD, existió una clasificación al nivel del estado del arte de la categoría *happy* [1], con una precisión media de 92.39 % y máxima de 95.83 %. Utilizando la métodología PCA y el algoritmo LDA, se obtuvo un desempeño moderado en la clasificación de la categoría *angry*, con una precisión media de 60.98 % y máxima de 71.72 %. Aunque los resultados de esta categoría puedan parecer bajos, hay que tomar cuenta que el modelo no repitió participantes en su entrenamiento y prueba, por lo que sus evaluaciones siempre fueron con rostros nuevos, esto nos indica que se midió su capacidad de generalización, evitando resultados con sobreajuste.

El resto de categorías definidas (*neutral* y *sad*) presentaron un rendimiento bajo con estas metodologías, pues sus resultados individuales fueron muy cercanos o menores a una clasificación aleatoria (33 %) dada la condición de correcta clasificación de la clase *happy*. Esto puede ser por múltiples causas: una posibilidad es que las expresiones de estas dos categorías no fueron posadas correctamente por

muchos participantes, otra es que la similitud de estas expresiones las hace más difícil de diferenciar por lo que se requieren métodos mas sofisticados para separarlas. También es necesario considerar que las condiciones de iluminación pudieron no ser consistentes a través del conjunto, dificultando más la clasificación de esas categorías.

Tabla 1: Desempeño de los algoritmos de clasificación sobre la categoría *happy* del conjunto utilizando los tres métodos de extracción de características

Método de extracción de características	Método de clasificación	Precisión (%)			Mejor
		Promedio	Desviación relativa	Máxima	
PCA	LDA	79.93	13.20	93.57	
	QDA	75.03	13.78	84.95	
	DT	55.45	8.31	61.37	
	SVM_{Lineal}	83.73	12.96	95.83	⋆
	SVM_{RBF}	74.23	9.50	85.30	
LMD	LDA	90.00	4.04	93.30	
	QDA	90.06	5.03	95.50	⋆
	DT	81.00	9.23	87.06	
	SVM_{Lineal}	79.65	6.62	86.47	
	SVM_{RBF}	84.94	7.24	62.46	
PCA+LMD	LDA	88.12	6.22	92.69	
	QDA	92.39	5.80	96.83	⋆
	DT	81.35	5.10	86.10	

Tabla 2: Desempeño de los algoritmos de clasificación sobre la categoría *angry* utilizando dos métodos de extracción de características

Método de extracción de características	Método de clasificación	Precisión (%)			Mejor
		Promedio	Desviación relativa	Máxima	
PCA	LDA	60.98	11.41	71.72	⋆
	QDA	60.49	13.30	70.71	
	SVM_{Lineal}	59.84	18.22	75.00	
	SVM_{RBF}	59.84	18.22	75.00	
LMD	LDA	58.95	12.93	69.17	⋆
	QDA	58.65	9.33	65.33	
	SVM_{RBF}	55.46	10.13	62.46	

4. Conclusiones

Tras analizar la congruencia de los resultados de la clasificación, se llegó al entendimiento que las herramientas desarrolladas en el estudio fueron efectivas para la captura, preprocesamiento, compilación, extracción de características y clasificación de los datos. Los resultados obtenidos de detección de expresiones faciales utilizando la distancia de puntos de referencia con el algoritmo QDA en el conjunto de datos fueron muy buenos para la categoría *angry* y *happy*. Es una meta razonable para subsiguiente investigación desarrollar métodos de extracción y clasificación que permitan mayor éxito en la separación de la categoría *neutral* y *sad*.

La decente cantidad de datos obtenidos en el estudio (9000 imágenes) demostró que la elaboración de experimentos a distancia es una opción viable para alcanzar a un considerable número de participantes, incluso en medio de una crisis de salud como la pandemia por COVID-19. Por otro lado, la existencia de un esfuerzo inicial para la construcción de un dataset como este es un camino que podría permitir aumentar la precisión de otras aplicaciones ya desplegadas o por implementarse en el país. Aún así, existen varías vías sobre las cuales subsiguiente investigación puede ayudar con la mejora continua de este, construyendo un conjunto más robusto y con mayor utilidad, como lo son: conseguir una mayor cantidad de participantes, cambiar la forma en la que se dan las instrucciones, agregar otra sesión de ensayos por cada voluntario, y comparar las mediciones con otros parámetros capturados de forma simultánea, e.g. la medición de la frecuencia cardíaca.

Declaración de conflictos de interés

El autor declara no tener ningún conflicto de interés para este trabajo.

Agradecimientos

Se agradece infinitamente la colaboración de los 75 participantes del experimento reclutados por redes sociales y también a las personas que apoyaron en su difusión.

Referencias

[1] M. Revina and W. S. Emmanuel, "A survey on human face expression recognition techniques," 2018.

[2] N. Samadiani, G. Huang, B. Cai, W. Luo, C.-H. Chi, Y. Xiang, and J. He, "A review on automatic facial expression recognition systems assisted by multimodal sensor data," *Sensors*, 2019.

[3] M. Conley, D. Vellarco, E. Rubien-Thomas, A. O. Cohen, N. Tottenham, and B. Casey, "The racially diverse affective expression (radiate) face stimulus set," *Psychiatry Research*, Dec 2018.

[4] K. M. Prkachin, "Assessing pain by facial expression: Facial expression as nexus," *Pain Res Manag*, 2009.

[5] M. Mangaiyarkarasi and A. Geetha, "Cursor using face expressions for human-computer interaction," 2014.

[6] M. Nishigaki and T. Shirakata, "Driver attention level estimation using driver model identification," 2019.

[7] S. Jerrita and M. Murugappan, "Physiological signals based human emotion recognition: a review," 2011.

[8] P. R. McDonald, A. M. Slater, and C. A. Longmore, "Covert detection of attractiveness among the neurologically intact: Evidence from skin-conductance responses," 2008.

[9] C. Xiefeng, Y. Wang, S. Dai, and P. Zhao, "Heart sound signals can be used for emotion recognition," 2019.

[10] G. O. Ganfure, "Using video stream for continuous monitoring of breathing rate for general setting," 2019.

[11] A. Koakowska, A. Landowska, and M. Szwoch, "Emotion recognition and its applications," Jan 1970.

[12] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[13] A. L. Ramadhani, P. Musa, and E. P. Wibowo, "Human face recognition application using pca and eigenface approach," *Second International Conference on Informatics and Computing (ICIC)*, 2017.

[14] S. Kuhanesan and S. Thuseethan, "Eigenface based recognition of emotion variant faces," *SSRN*, 2016.

[15] F. Zahra Salmam, A. Madani, and M. Kissi, "Facial expression recognition using decision trees," 2016.

[16] C. Yu, L. Jinxu, and Z. Fudong, "Comparative study on face recognition, svm of one-against-one and one-against-rest methods," *International Conference on Future Generation Communication and Networking*, 2014.