Clasificación eficiente usando la Característica de Euler

Erik Amézquita [1] Mario Canul [2] Antonio Rieser [3]
erik.amezquita@cimat.mx

[1]DEMAT, UGto

[2]CIMAT

[3]CONACYT-CIMAT

28 de septiembre de 2017

Vistazo general

Pregunta, Problema y Objetivo

- **¿Puede la topología decirnos algo de estas máscaras?**
- Clasificación eficiente de objetos no sujeta a subjetividades del usuario.
- Establecer criterios de clasificación basados en características geométricas y topológicas del objeto.
- Usar la idea de **gráfica de característica de Euler (CE)** como sugirieron Richardson y Weirman en el 2014 en [5].

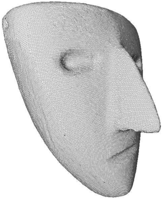
Característica de Euler (CE)

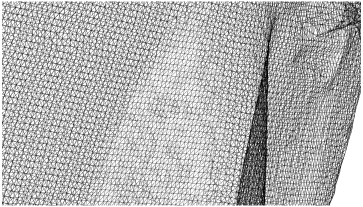
Consideremos un objecto n dimensional $X = (V_0, V_1, \dots, V_n)$ y su característica de Euler (CE):

$$\chi = \sum_{k=0}^{n} (-1)^k \; |V_k|$$

No. de celdas k dimensionales

La CE es un **invariante** topológico del objeto.





La filtración

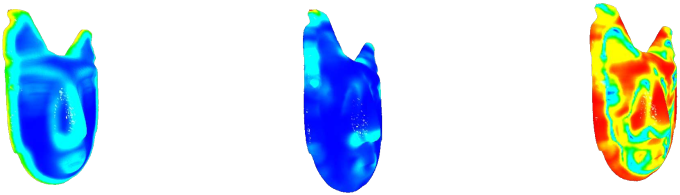
Fijamos una función de filtración g para los vértices y luego la extendemos al resto de las k-celdas:

$$g_k(\ \{v_0, v_1, \ldots, v_k\}\ ) = \min_{0 \le i \le k} \{\ g(v_i)\ \}$$

$g_k : V_k \to [a, b]$

una k-celda

Una función $g : V_0 \to [a, b]$
V_0 el conjunto de vértices; $[a, b]$ intervalo fijo.



Umbralización

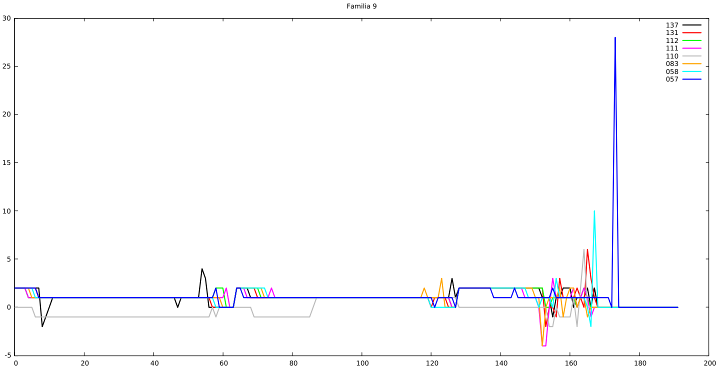
El intervalo $[a, b]$ es dividido en T umbrales equiespaciados $a = t_0 < t_1 < t_2 < \ldots < t_T = b$. Consideramos la CE en el i-ésimo intervalo:

$$\chi_i = \sum_{k=0}^{n} (-1)^k \mid V_k^{(i)} \mid.$$

No. de k celdas c_k tales que $g_k(c_k) \geq t_i$

La GCE

La **gráfica de característica de Euler (GCE)** es simplemente comparar χ_i vs t_i



Algorítmicamente hablando I

Los valores numéricos $g(v)$ ya están calculados para todo vértice v.

Input: $g,\ T$
$\chi[T] \leftarrow 0$ ▷ Valores χ_i
for all $k = 1 \to n$ **do** ▷ dimensiones
  $H[T] \leftarrow 0$ ▷ histograma
  for all $i = 1 \to N_k$ **do** ▷ k-celdas
    $g_k \leftarrow \min g$
    $b \leftarrow bin(g_k)$ ▷ $\lfloor g_k \times T \rfloor$
    $H[b] = H[b] + 1$
  $c \leftarrow 0$
  for all $i = T \to 1$ **do** ▷ umbrales
    $c \leftarrow c + H[i]$
    $\chi[i] \to \chi[i] + (-1)^k c$
return χ

Algoritmicamente hablando II

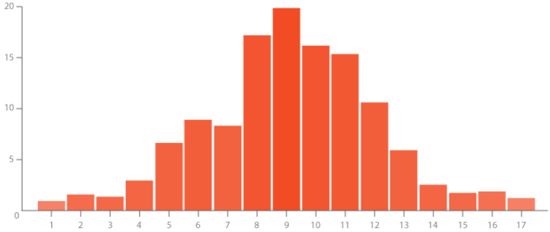
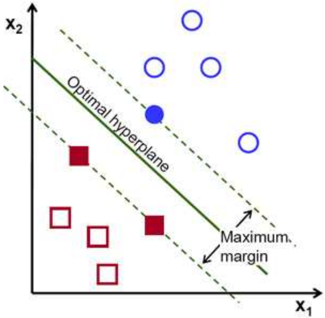


Figura: El algoritmo es eficiente al computar los valores del histograma $H[T]$ una única vez

El algoritmo tiene complejidad $O(N(V+T)) \approx O(V)$, V número de vértices.

Support Vector Machine (SVM)

- Método supervisado: conjunto de entrenamiento y conjunto de prueba.
- Caso separable binario: puntos $\vec{x}_i \in \mathbb{R}^n$ que pertenecen a clase $y_i \in \{1, -1\}$.
- Dividas por el hiperplano $\langle \vec{w}, \vec{x} \rangle + b = 0$.



Entrenamiento

- Se cumplen las condiciones

$$\langle \vec{w}, \vec{x}_i \rangle + b \geq 1 \text{ para } y_i = +1, \tag{1a}$$

$$\langle \vec{w}, \vec{x}_i \rangle + b \leq 1 \text{ para } y_i = -1. \tag{1b}$$

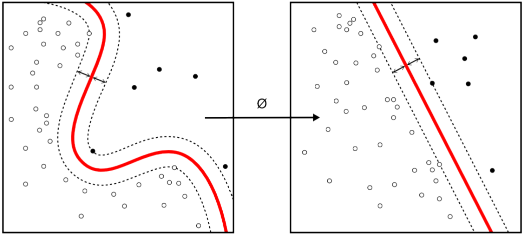
- Ello se combina como

$$y_i(\langle \vec{w}, \vec{x}_i \rangle + b) - 1 \geq 0 \quad \forall i. \tag{2}$$

- Los vectores de soporte son aquellos donde se da la igualdad.
- Éstos definen hiperplanos H_1, H_2.
- La distancia entre éstos es $\frac{1}{||\vec{w}||}$.
- Minimizar $||w||$ dada la restricción (2)

Prueba y Kernel

- Dado un punto $\vec{x}$, su clase es $sgn(\langle \vec{w}, \vec{x} \rangle + b)$.
- $\Phi : \mathbb{R}^n \to \mathcal{H}$ espacio de Hilbert.
- $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}, \quad (\vec{x}, \vec{y}) \mapsto \langle \Phi(\vec{x}), \Phi(\vec{y}) \rangle_{\mathcal{H}}$.
- $K(\vec{x}, \vec{y}) = (\langle \vec{x}, \vec{y} \rangle + 1)^p$ da un clasificador polinomial de grado p.



Análisis de datos

El principal problema afrontado fue dar una nueva clasificación al conjunto de 128 máscaras digitalizadas por el Instituto Nacional de Antropología e Historia (INAH.) Acorde a la clasificación de máscaras manejada por el INAH, las 128 se dividen en 9 familias distintas, por lo que se buscó dar una clasificación en 9 grupos.

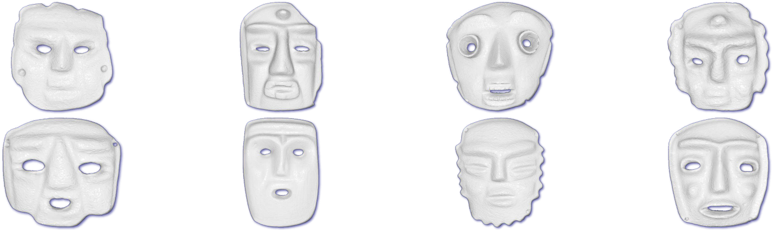
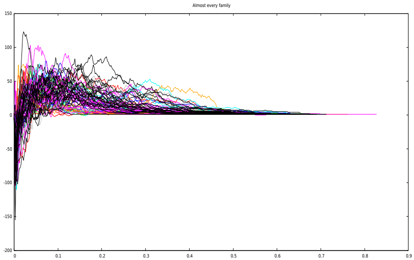


Figura: Familia 2 en la clasificación **original**

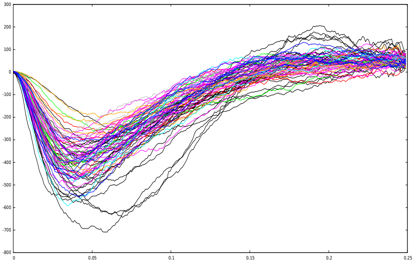
Figura: Muestra de la familia 9 en la clasificación **original**

Primeras GCEs

Al usar curvaturas como filtraciones, las GCEs asociadas de las máscaras no muestran patrones claros.



(a) Curvatura media



(b) Índice de Forma

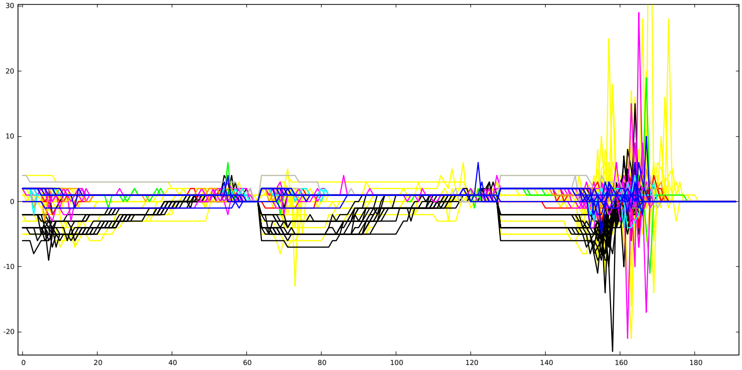
Figura: Gráficas de CE para curvatura media e Índice de Forma en $T = 256$ umbrales. Cada una de las 9 familias originales fue trazada con un color distinto

Las proyecciones como filtración

Aprovechamos que cada máscara está encajada en el cubo $[-1,1]^3$ con centro de masa en el origen. Las filtraciones fueron las distancias de cada vértice a los planos $x = 1,\ y = 1,\ z = 1$.

(a) Horizontal (b) Vertical (c) Frontal

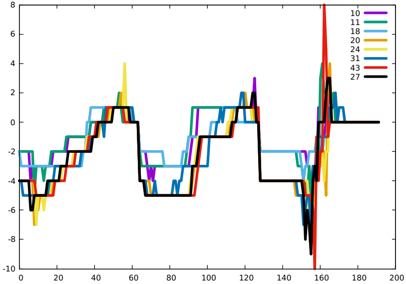
La GCE a partir de la concatenación de las **tres proyecciones principales** con 64 umbrales por proyección proveé de un mejor prospecto para obtener una clasificación coherente de objetos.

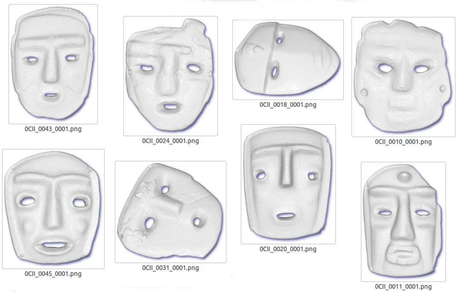


De las GCEs al SVM

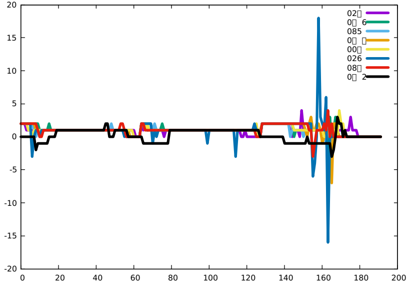
- El SVM usó la mitad del conjunto de máscaras como entrenamiento y el resto como prueba. Se obtuvo una nueva clasificación.
- El número de especímenes por familia es más homogéneo. Ahora únicamente dos de las nueve familias contiene menos de 10 representantes.
- De los siete grupos restantes, se eligieron 8 representantes de cada una y se graficaron sus GCEs.
- Colores distintos se refieren a items distintos.

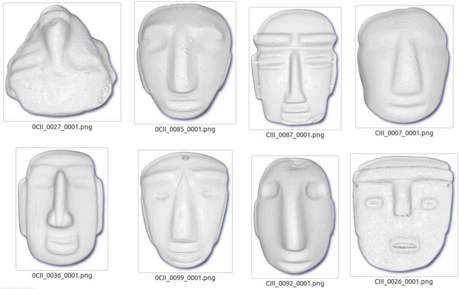
Familia 2 (clasificación nueva)



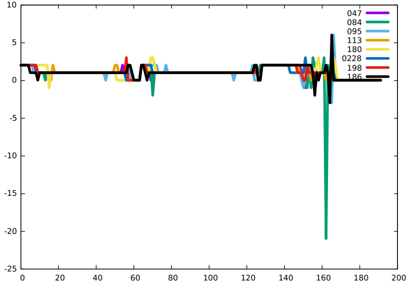


Familia 3 (clasificación nueva)

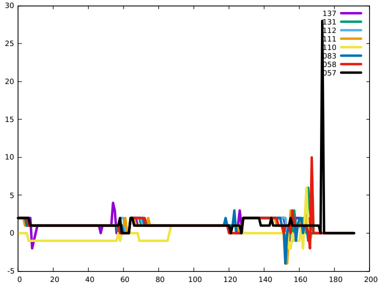




Familia 5 (clasificación nueva)



Familia 9 (clasificación nueva)



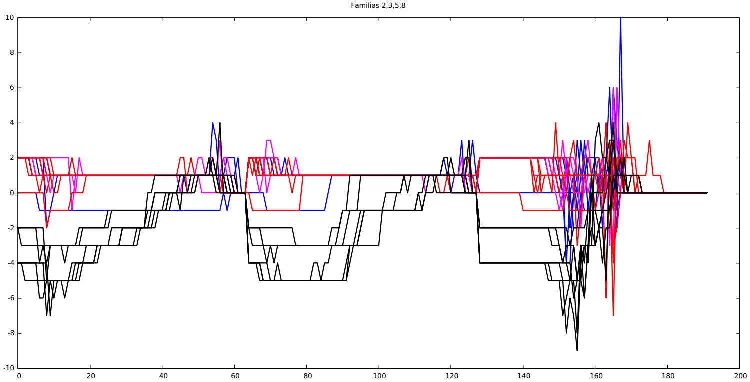


Figura: GCEs de las cuatro familias previas después de remover *outliers*.

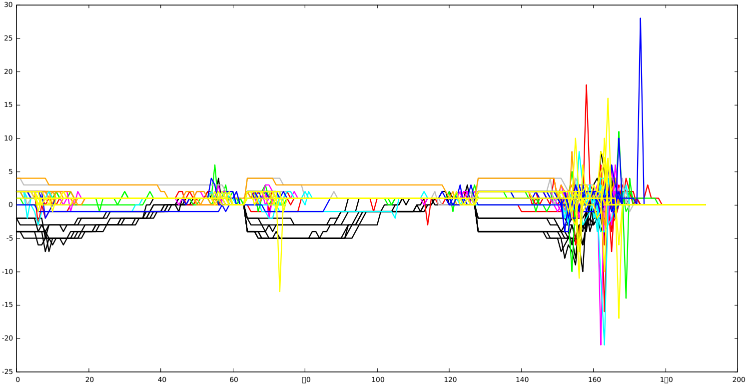


Figura: GCEs tomando 8 miembros de cada familia acorde a la nueva clasificación, eliminando aquellos cuya GCE sobresalga del resto.

De las GCEs al USVM

El caso más fácil sí funciona:

Figura: Familia 2 en la clasificación **original**

Figura: Familia 9 en la clasificación **original**

Resultado esperado

```
Optimal vector w:
comparing 11 vs 19
Margin: 0.2332
Vector 1: -
Vector 2: -
Vector 3: -
Vector 4: -
Vector 5: -
Vector 6: +
Vector 7: -
Vector 8: -
Vector 9: -
Vector 10: -
Vector 11: -
Vector 12: +
Vector 13: -
Vector 14: +
Vector 15: +
Vector 16: +
Vector 17: +
Vector 18: +
Vector 19: +
Vector 20: +
FAMILY 2 vs FAMILY 9 using T =
```

Con otras familias se porta extraño

Figura: Muestra de la familia 3 en la clasificación **original**

Figura: Muestra de la familia 4 en la clasificación **original**

Una a una vamos excluyendo

Figura: Exclusión ordenada

Un ejemplo más

Figura: Familia 5 en la clasificación **original**

Figura: Muestra de la familia 4 en la clasificación **original**

Una a una vamos excluyendo

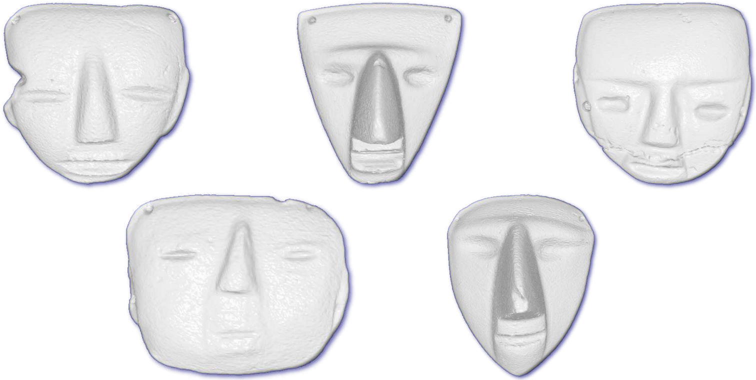


Figura: Exclusión ordenada

Comentarios Finales

- El cómputo de la GCE es una operación sencilla de complejidad y memoria lineales. Procesa rápidamente objetos de decenas de miles de vértices.
- Debido a su rapidez, este algoritmo hace pensar en aplicaciones en tiempo real de reconocimiento de patrones de superficies y objetos en general, no necesariamente piezas arqueológicas.
- Una mayor cantidad de máscaras puede proveer de mejores conjuntos de entrenamientos y por ende, mejores clasificaciones.
- Más especímenes permitirán también experimentar con métodos de clasificación no supervisada.

Referencias

- M.A. Armstrong *Basic Topology* Springer-Verlag (1983).
- C. Burges "A Tutorial on Support Vector Machines for Pattern Recognition". *Data Mining and Knowledge Discovery* Vol.2 pp.121-167, 1998.
- Z. Karnin et. al., "Unsupervised SVMs: On the Complexity of the Furthest Hyperplane Problem." *JMLR: Workshop and Conference Proceedings* Vol. 23 (2012) 2.1 - 2.17
- V. Paulsen, M. Raghupathi, *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces* Cambridge University Press (2016).
- E. Richardson, M. Weirman, "Efficient classification using the Euler Characteristic". *Pattern Recognition Letters* Vol.49, pp.99-106, 2014.

Agradecimientos

IPICYT

Instituto Nacional de Antropología e Historia

RED TEMÁTICA
TECNOLOGÍAS DIGITALES
PARA LA DIFUSIÓN DEL
PATRIMONIO CULTURAL

CONACYT
Consejo Nacional de Ciencia y Tecnología