Estudio de conjuntos muestrales usados en clasificación automática de imágenes

Adrián Carballal[a] (adrian.carballal@udc.es), Juan Romero[a] (jj@udc.es)
Penousal Machado[b] (machado@dei.uc.pt), Maria Luz Castro Pena[a] (maria.luz.castro@udc.es)

[a]Facultade de Informática, Universidade da Coruña – Campus de Elviña, 15071 A Coruña, España
[b]Department of Informatics Engineering, Universidade de Coimbra – Polo II, 3030-290 Coimbra, Portugal

***Abstract* — En los últimos años se han realizado varios trabajos en los que se ha intentado abordar la clasificación automática de imágenes. En dichas aproximaciones se han utilizado distintos conjuntos de imágenes de muestra provenientes de diferentes portales web de fotografía pero no se ha estudiado la capacidad de generalización que ofrecen cada uno de ellos. En nuestro análisis se muestran algunos de estos conjuntos de imágenes experimentales más utilizados en el campo de la estética computacional. Se realizan una serie de experimentos por medio de clasificadores binarios y características simples cruzando dichos conjuntos muestrales. En este artículo se estudia si resulta trivial la elección de las imágenes atendiendo a la capacidad de generalización obtenida por cada uno de los clasificadores empleados.**

***Index Terms* — Estética, validación cruzada, entropía.**

I. INTRODUCCIÓN

En la actualidad existen algunos trabajos en los que se ha intentado clasificar automáticamente diferentes conjuntos de imágenes o datasets siguiendo distintos criterios: estética, originalidad, temática, etc. En todos estos trabajos hay un elemento clave: las imágenes utilizadas como referencia. Entendemos que dichos conjuntos de imágenes son un componente crítico para poder obtener resultados concluyentes y fácilmente comparables. Hasta la fecha no se ha estudiado si los conjuntos muestrales utilizados son capaces de proporcionar una representación generalizada para dichas tareas de clasificación. En este artículo estudiaremos algunos previamente usados en experimentos computacionales relacionados con la calidad estética.

Debe quedar claro que en dichos trabajos se habla de clasificar imágenes atendiendo a criterios meramente estéticos. Esto conlleva a atender a aquellas propiedades principalmente relacionadas con la belleza estética como la forma, y obviando en la medida de lo posible aquellas importantes desde un punto de vista artístico como el contenido. Ambos conceptos, arte y estética, están fuertemente relacionados, pero, mientras el contenido se suele valorar de forma subjetiva, la forma en si misma de una imagen puede valorarse objetiva y/o cuantitativamente mediante distintas características. Las características usadas tanto en el estado del arte como en este trabajo son generales por lo que pueden ser calculadas para cada imagen independientemente de su naturaleza y contenido.

A continuación se enumeran distintos trabajos, cada uno de los cuales presentan distintas características que pretenden ser útiles a la hora de medir la calidad estética de una fotografía.

II. ESTADO DEL ARTE

Dentro del grupo de trabajos relacionados con la clasificación estética automatizada podemos destacar aquellos llevados a cabo por Datta et al. [1], Wong et al. [2], Ke et al. [3] o Luo et al. [4]. En estos trabajos los autores han propuesto distintas características basadas en componentes técnicas como luminosidad, saturación, Regla de los Tercios, etc., en busca de aquellas que mejor se adapten para dicha tarea. En todos los casos han realizado sus experimentos usando fotografías obtenidas de distintos portales web y las votaciones que sus usuarios realizan sobre las mismas.

Si bien este enfoque proporciona grandes conjuntos de datos creados por un tercero, lo que debería reducir al mínimo las posibilidades de que sea parcial, tiene varios defectos. El sistema de votación empleado no es tan controlado como en un test psicológico, y varios factores exógenos pueden influir en las puntuaciones de la imagen. No es posible tener toda la información acerca de las personas y las circunstancias en las que participaron. También es difícil saber lo que los usuarios están evaluando cuando votan.

Por ejemplo, Datta et al. [1] utilizan un conjunto de imágenes obtenidas de "Photo.net". En dicho portal los usuarios pueden puntuar cada imagen de acuerdo a su "estética" y "originalidad". Sin embargo, estas puntuaciones están muy correlacionadas [1], lo que indica que los usuarios no son capaces de distinguir entre ambos criterios.

En el caso de Ke et al. [3] utilizan imágenes pertenecientes a "DPChallenge.com". También se trata de un portal fotográfico pero en él se incluye el hándicap de que es una competición, por lo que las posibilidades de votos sesgados son aún mayores.

Al tratarse de websites fotográficos, los usuarios prefieren con frecuencia algunas imágenes con determinados estilos, que pueden constituir un cierto sesgo en el conjunto de datos. Algunos de los participantes en estos sitios web son, de hecho, fotógrafos profesionales. Por otra parte, la selección de imágenes no está bajo control, la valoración estética puede ser muy influenciada por la semántica de los contenidos, la novedad, originalidad, etc.

Los conjuntos de imágenes presentados por Datta et al. [1] y de Ke et al. [3] han sido utilizados como referencia por otros autores comparando nuevas características pero existe un problema intrínseco en el enfoque empleado. El hecho de que determinadas características sean útiles a la hora de clasificar un conjunto de imágenes concreto no implica que su valía pueda ser considerada como universal o sus resultados como generalizables. Para poder llegar a unos resultados fiables sería necesario utilizar como referencia conjuntos de imágenes lo suficientemente generales como para que al usar otras distintas pudiésemos estar relativamente seguros de la clasificación obtenida.

La metodología empleada en los trabajos comentados tiene, a nuestro entender, un problema subyacente: se entrenan y validan distintos clasificadores empleando el mismo conjunto o conjuntos de imágenes obtenidos de la misma fuente. La utilización de dicha metodología sin haber estudiado los conjuntos muestrales en profundidad previamente puede incidir en la obtención de resultados que no garanticen una correcta universalidad. Por esta razón, creemos que la utilización de distintos conjuntos de imágenes durante la fase de aprendizaje puede ayudar a evaluar la consistencia y la coherencia de los resultados obtenidos.

En este trabajo se realizan dos experimentos con unos clasificadores simples y empleando características básicas relacionadas con valores estadísticos de la distribución de la intensidad y de estimadores de la entropía. Emplearemos en la fase de entrenamiento y validación dos enfoques: i) utilizar el mismo conjunto de imágenes en ambas fases de aprendizaje y ii) emplear distintos conjuntos de imágenes. Como conjuntos de imágenes muestrales se utilizarán aquellos diseñados por Datta et al. [1] y por Ke et al. [3]. Con ello se pretende mostrar como la deficiencia de universalidad comentada está patente entre ambos conjuntos.

Este artículo se estructura de la siguiente manera: i) se muestran distintos dataset relativos a trabajos sobre clasificación estética pertenecientes al estado del arte, ii) se presenta un pequeño conjunto de características que utilizaremos para catalogar imágenes atendiendo a su calidad estética, iii) se enseña el método utilizado para realizar la clasificación, iv) se presentan los resultados obtenidos, tanto en clasificación individual de cada uno de los datasets utilizados como cruzados para comprobar la generalización alcanzada, v) se discutirán los resultados obtenidos.

III. DATASETS SOBRE ESTÉTICA

En la sección anterior hemos comentado la existencia de diversos trabajos relacionados con la clasificación estética de imágenes. En este apartado se detallan algunos de los conjuntos de imágenes más empleados en este tipo de tareas:

A. Imágenes de Photo.net (2006)

Datta et al. [1] crearon un dataset a partir del portal fotográfico "Photo.net", el cual ha sido utilizado posteriormente en experimentos de clasificación estética de imágenes.

Este portal contiene cerca de un millón de imágenes pertenecientes a más de 400,000 usuarios. Cada fotografía puede ser comentada y puntuada siguiendo dos criterios, estética y originalidad. Toda esta información es pública para cualquier usuario. Una imagen puede recibir puntuaciones comprendidas entre 1 y 7, siendo 1 la peor valoración posible y 7 la mejor.

Los creadores de dicho dataset han publicado vía Internet un fichero en el cual se especifican los identificadores de las imágenes que lo componen, las estadísticas individualizadas de los votos obtenidos por cada imagen dentro del rango tanto en el caso de originalidad como en el de estética, así como los valores de las características que presentan. El dataset completo está formado por 3,581 imágenes que han sido valoradas por al menos 2 personas diferentes. A nuestro entender, aunque este dataset haya sido utilizado posteriormente en diversos experimentos, tiene varios problemas asociados: (i) el conjunto asemeja ser demasiado pequeño y no dan ninguna razón sobre la elección del tamaño, (ii) asumen como representativo el valor medio asociado a una imagen, el cual ha sido obtenido en algunos casos a partir de solo dos votos.

En los experimentos de clasificación que se ha utilizado este conjunto de imágenes normalmente se realiza una distribución previa con el fin de obtener dos conjuntos disjuntos de imágenes de alta y baja calidad estética [1][2][5]. Dichos conjuntos se determinan mediante las valoraciones estéticas que han hecho los usuarios por

medio de sus puntuaciones medias. Datta et al. [1] denominan a cada uno de estos dos conjuntos como Low (baja calidad) y High (alta calidad) respectivamente.

Fig. 1. Ejemplos de imágenes pertenecientes al dataset utilizado por Datta et al. [1] y Wong et al. [2] obtenidas del portal "Photo.net".

La división de dichos conjuntos se realiza de la siguiente manera: i) aquellas imágenes con una puntuación media ≥ 5.8 son catalogadas dentro del conjunto High; ii) toda fotografía con una puntuación media ≤ 4.2 es catalogada dentro del conjunto Low. Finalmente, los conjuntos experimentales empleados por Datta están compuestos por un total de 832 y 760 imágenes para los conjuntos High y Low respectivamente. En la Fig. 1 se muestran algunos ejemplos de imágenes de "Photo.net" que pertenecen a este conjunto.

Datta et al. [1] extraen algunas características visuales basadas en la intuición, que según ellos pueden ser capaces de distinguir entre imágenes estéticamente agradables y desagradables. Intentan explorar la relación entre las emociones que las imágenes evocan en la gente. Entre las características utilizadas se encuentran: exposición a la luz, saturación, Regla de los Tercios, composición de regiones, convexidad de forma, etc. Con dicho conjunto de métricas alcanzaron un 70,12% de acierto en la clasificación global utilizando Maquinas de Soporte Vectorial o SVM, que se corresponde con un 68,08% en el caso de las imágenes pertenecientes al conjunto High y un 72,31% en el caso de las imágenes del conjunto Low.

Wong et al. [2] exponen un total de 44 métricas agrupabas en tres categorías de características globales: (i) basadas en técnicas básicas como nitidez y contraste, (ii) normas fotográficas y (iii) ajustes de cámara. En su caso utilizan un subconjunto de las imágenes del experimento original, 3,161 imágenes de las 3,581 iniciales. Argumentan que les ha sido imposible disponer del conjunto de entrenamiento completo debido a la eliminación de algunas de estas fotografías del portal "Photo.net". La clasificación la realizan mediante SVM con un kernel lineal. Utilizan una validación cruzada mediante 5 ejecuciones independientes sobre los mismos datos muestrales. Son capaces de clasificar correctamente el 78.2% de las imágenes, correspondiente con el 82.9% para aquellas consideradas de alta calidad estética o High frente al 75.6% de las de baja calidad estética o Low.

B. Imágenes de CUHK (2006)

Otro trabajo centrado en la clasificación estética en el cual se crea un dataset a partir de imágenes pertenecientes a un portal de fotografía es el expuesto por Ke et al. [3]. Sobre dicho trabajo se han realizado experimentos posteriormente, entre los que destaca el de Luo et al. [4] por ser el que ha ofrecido los mejores resultados hasta la fecha.

Las imágenes de las que consta este dataset se obtuvieron del portal de fotografía "DPChallenge.com". Todas las imágenes que aparecen en este portal han sido puntuadas por alguno de sus usuarios. Dichas puntuaciones se encuentran en el rango [1, 10], siendo 1 la peor valoración y 10 mejor valoración posible. El dataset original se compone de un total de 60,000 imágenes. Cada una de ellas ha sido valorada por al menos 100 usuarios, aunque no se especifica en ningún caso el promedio de votaciones por imagen y su desviación media.

Al igual que con el dataset anterior, a la hora de realizar experimentos de clasificación estética se han creado dos conjuntos. Del total de imágenes se extrajeron el 10% superior e inferior de las fotografías una vez ordenadas previamente por su valoración media. El 10% superior se cataloga como de alta calidad estética y el 10% inferior como de baja, por lo que cada uno de los dos subconjuntos tiene un total de 6,000 imágenes.

Una vez separadas las fotografías de alta calidad de las de baja calidad estética, Ke et al. [3] las subdividen en dos nuevos conjuntos de igual tamaño determinados aleatoriamente. De este modo, se obtienen finalmente 4 conjuntos de 3,000 imágenes cada uno, dos formados por imágenes de baja calidad y otros dos de alta calidad estética. Cada uno de estos conjuntos es utilizado para propósitos distintos, uno de cada tipo para entrenar los sistemas propuestos y los otros dos para validar su capacidad. En la Fig. 2 se muestran algunos ejemplos de imágenes pertenecientes a este conjunto.

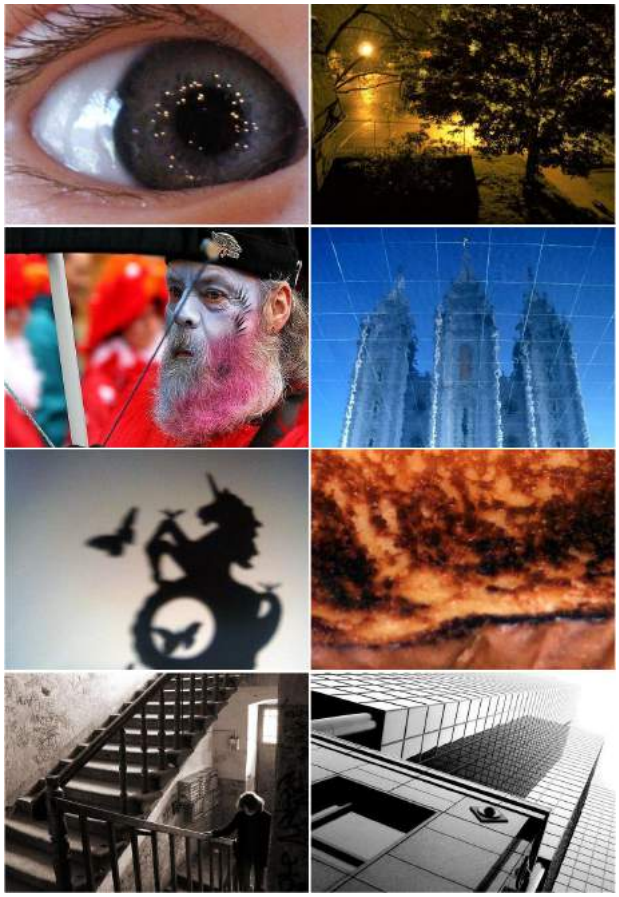

Fig. 2. Ejemplos de imágenes pertenecientes al dataset *CUHK* [3][4] obtenidas del portal "DPChallenge.com".

El mejor resultado obtenido con este dataset ha sido el de Luo et al. [4] con un 93% de acierto utilizando características basadas en claridad, luminosidad, simplicidad, composición geométrica, armonía de color y extracción de regiones (identificaron del fondo y primer plano). En cambio, en el experimento original, Ke et al. [3] obtuvieron una tasa acierto del 72%, utilizando medidas relacionadas con la distribución espacial de los ejes, distribución del color, matiz, y borrosidad.

C. Otros datasets publicados (2008)

En el año 2008, como parte del artículo de Datta et al. [6], se proponen a la comunidad científica cuatro datasets de prueba para la realización de experimentos relacionados con la predicción de resultados estéticos, predicción de clase estética y predicción de la emoción. Todos estos datasets se componen de imágenes obtenidas de portales de fotografía donde los usuarios pueden realizar una valoración de las mismas. Entre dichos portales se encuentran "Photo.net", "DPChallenge.com", "Terragalleria.com" y "Alipr.com". En el caso de los dos primeros se tratan de datasets diferentes a los comentados anteriormente.

El dataset obtenido de "Photo.net" está formado por 20,278 imágenes, de las cuales se encuentra disponible sus respectivas votaciones. Algunas de las imágenes ya no se pueden recuperar del portal de fotografía al no estar disponibles actualmente. La media de valoraciones por imagen es de 12 votos con una desviación típica de 13.

En el caso de las imágenes relativas al portal "DPChallenge.com", se trata de un total de 16,509 imágenes, cada una de las cuales está votada al menos por un usuario. La media de valoraciones por imagen es de 205 con una desviación de 53.

El tercer conjunto de imágenes mostrado por Datta et al. [6] está compuesto de 14,449 imágenes obtenidas de "Terragalleria.com". Dicho portal se compone de fotografías de viajes de Quang Tuan-Luong, y es considerada como una de las colecciones fotográficas más importantes del US National Park. Todas las fotografías han sido tomadas por él, lo cual lo diferencia de los otros datasets vistos anteriormente, pero siguen siendo los usuarios los que valoran la calidad de las fotografías en una escala de 1 a 10. La media de valoraciones por imagen es de 22 con una desviación media de 23.

Por último, del buscador y etiquetador "Alipr.com" se han recolectado un total de 13,010 imágenes, algunas de ellas repetidas. Dichas imágenes han sido valoradas por los usuarios atendiendo a 10 emociones. Al igual que sucede con las imágenes de CUHK, no se posee información estadística relativa a las valoraciones realizadas. No tenemos constancia de que ninguno de estos conjuntos de imágenes haya sido utilizado en experimentos de clasificación computacional.

IV. Experimento de validación cruzada entre distintos datasets

Como hemos comentado, han sido publicados con anterioridad conjuntos de imágenes diseñados para

estudiar esta temática. De los que hemos mencionado, sólo se han utilizado experimentalmente el de *Photo.net* [1] y el *CUHK* [3]. Por este motivo, vamos a realizar un experimento simple que nos permita comprobar si por medio de esos conjuntos de imágenes se podrían clasificar estéticamente otras fotografías aunque no fuesen obtenidas de la misma fuente.

Utilizaremos un clasificador binario alimentado por un conjunto de características básicas (estadísticas y estimadores de la entropía). En nuestro caso entrenaremos dos clasificadores, uno utilizando las imágenes de *Photo.net* y otro las de *CUHK*. Se comprobará la capacidad de clasificación de ambos de dos formas distintas, utilizando el mismo conjunto de imágenes empleado en el entrenamiento para su validación y empleando otro distinto. En el caso del conjunto de imágenes de *CUHK*, ha sido posible disponer del conjunto completo de imágenes empleado por Ke et al. [3]. Sin embargo, ha sido imposible recuperar el conjunto completo de imágenes detallado por Datta et al. [1], formado por 3,581 imágenes. En concreto, sólo se han podido recuperar un total de 3,247 imágenes debido a que algunas ya no están disponibles en "Photo.net". Este mismo problema, como ya hemos comentado, también les ha sucedido a otros investigadores [2].

A continuación especificaremos las métricas estadísticas y las relativas a la entropía empleadas.

V. Características básicas utilizadas

La mayoría de los experimentos en el área de la computación realizados sobre estética siguen el mismo patrón. Se utilizan distintos clasificadores computacionales, normalmente binarios, alimentados por un conjunto de datos que representan valores asociados a las imágenes muestrales. Por lo general se utilizan valores relacionados con aspectos técnicos como la luminosidad y la saturación, con el fin de encontrar aquellos que puedan obtener los mejores resultados posibles en tareas de clasificación de imágenes atendiendo a criterios estéticos.

La calidad de mejora que la elección de los valores pueda proporcionar no es la máxima de este trabajo. Estamos más interesados en comprobar si resulta trivial la elección de las imágenes para realizar experimentos de esta índole. Es por ello que vamos a utilizar en nuestros experimentos dos conjuntos de valores, uno obtenido a partir de la media y la desviación típica del valor de los pixeles que componen cada imagen y el otro usando valores que tratan de estimar la entropía relativa de una imagen a partir de la diferencia entre el valor de un pixel con respecto a sus pixeles vecinos. A continuación explicamos ambos conjuntos.

Antes de calcular cada una de estas características se ha realizado una transformación de cada imagen. Primero se ha redimensionado a 256x256 pixeles y después se ha transformado a un modelo de color RGB con una profundidad de color de 8-bit por canal de color escalado en el rango [0,255]. Con esto se ha conseguido que todas las imágenes compartan dimensiones y formato. Por último, se ha pasado cada imagen a un modelo de color HSV, el cual será básico para el cálculo del primer conjunto de métricas empleado.

Algunos pasos de esta transformación, como es el caso del cambio de la relación de aspecto a 1:1, constituye una pérdida de información y una deformación de la imagen, pero en experimentos anteriores en otros dominios se comprobó que dicha transformación no afectaba a la habilidad de este tipo sistemas para realizar clasificaciones de esta índole [5, 7].

Dentro del primer conjunto se han empleado dos medidas estadísticas para el color de cada imagen, la media y la desviación estándar. Ambas se calculan a partir del valor de la intensidad de los pixeles de cada imagen sobre el modelo de color HSV de la misma, a excepción del canal H (Matiz o Hue). Dado que el canal H es circular, la media y la desviación estándar se calculan basándose en la norma y el ángulo de los valores de Hue. De esta manera obtenemos un nuevo valor relativo al matiz en el que se minimiza el problema del modelo HSV para colores con valores extremos en los canales V y S. Con todo esto, obtenemos un total de 10 valores estadísticos: i) 6 para la media divididos en grupos de 2 para cada canal, incluyendo la variante para Hue y ii) 4 para la desviación típica, uno para cada canal al igual que la media. A este conjunto de valores nos referiremos en adelante como *AvgStd*.

El segundo conjunto de valores utilizados se corresponden con estimadores de la entropía de una imagen. La entropía mide el grado de desorden existente en un sistema, independientemente de la naturaleza del mismo. Según Arnheim [8], "el orden es una condición necesaria para cualquier cosa que ha de entender la mente humana". Basándose en dicha afirmación, existen varios trabajos en este ámbito que relacionan la entropía o grado de complejidad de una imagen con la belleza estética asociada [9]-[10]. Existen numerosos estimadores de la entropía aunque para este experimento se ha optado por la utilización de la Ley de Zipf. Según dicha ley, el rango estadístico de un evento es inversamente proporcional al tamaño del evento.

La ley de Zipf se basa en la observación de los fenómenos generados por organismos auto-adaptativos, como los seres humanos. Es comúnmente conocido como *principio del mínimo esfuerzo*. Una vez que un fenómeno

o evento ha sido seleccionado para su estudio, se examina la contribución de cada caso concreto con respecto al todo y se determina su rango de importancia o predominio. Informalmente, los eventos más pequeños tienden a ocurrir con mayor frecuencia, mientras que los eventos más grandes tienden a ocurrir con menor frecuencia. Una variante dentro de la Ley de Zipf usa el tamaño del fenómeno en lugar de su rango, generando una distribución de tamaño-frecuencia o *size-frequency*. En este experimento utilizaremos esta última formulación.

El cálculo de las características referentes a la frecuencia de aparición Zipf se realiza obteniendo la diferencia entre el valor de un pixel con cada uno de sus vecinos y se cuenta el número total de ocurrencias de dicho valor de diferencia. Después se ordenarán esos valores en orden descendente según el número de ocurrencias y se representan en un eje cartesiano según su valor y frecuencia. A partir de dicha gráfica se utilizarán como estimadores de la entropía la pendiente (m) y la de correlación lineal (R2) de dicha línea de tendencia. Como sucedía con la media y las desviación típica, obtenemos esos dos valores para los tres canales del modelo de color HSV. En caso del canal H, también se obtienen dos nuevos valores extras considerando una distancia circular. De este modo hablamos de un conjunto formado por 8 valores o estimadores de la entropía a los que nos referiremos en adelante como *SizeFreq*.

VI. Modelos de clasificación empleados.

Es muy común en trabajos derivados de la clasificación de imágenes atendiendo a criterios estéticos la utilización de un tipo de clasificadores denominados SVM o Máquinas de Soporte Vectorial. Los algoritmos SVM pertenecen a la familia de los clasificadores lineales. Permiten realizar clasificaciones entre conjuntos de datos a partir del margen máximo que existe de separación entre ambos.

Una SVM representa los datos muestrales en una superficie de decisión. Dichos datos, que normalmente son no linealmente separables, son convertidos por medio de una función o kernel a un espacio de características de mayor dimensión. Una vez hecho esto, el sistema determina una frontera de decisión que separa los puntos de muestra en clases distintas. Esta función de frontera, que representa un hiperplano, permite distinguir que datos pertenecen a cada una de las diferentes clases.

Existe un gran número de aplicaciones que permiten el uso de este tipo de clasificadores de forma sencilla e intuitiva, tanto para expertos como para principiantes. Una de las más usadas es WEKA (Waikato Environment for Knowledge Analysis) [11], que será la utilizada en estos experimentos.

De los trabajos anteriormente comentados, [1]-[4], utilizan este tipo de clasificadores, tanto con los parámetros por defecto como algunos específicos determinados empíricamente. En los experimentos realizados se ha utilizado una función o kernel lineal proporcionado por el paquete LibSVM [12] con sus parámetros por defecto (γ=3.7, ν=0.5, ε=0.1, sin normalización de los datos de entrada).

VII. Experimentos realizados

Se han realizado dos experimentos con los mismos valores, los mismos clasificadores y los mismos parámetros de aprendizaje e intercambiando únicamente los datos de entrenamiento y validación. Esto nos permite obtener dos conjuntos de resultados que nos servirán para valorar la capacidad de generalización obtenida con ambos conjuntos de imágenes de muestra.

En el primer experimento se han utilizados los 10 valores que hemos identificado como *AvgStd* y los 8 de *SizeFreq* individual y conjuntamente con las imágenes pertenecientes a *Photo.net* y *CUHK* respectivamente. En ambos casos se ha utilizado para entrenar y validar el clasificador el conjunto completo de imágenes, siguiendo la metodología empleada por [1]-[2].

En nuestro caso se utiliza un procedimiento de validación llamado *5-fold cross-validation*, que consiste en dividir el conjunto de patrones en 5 conjuntos disjuntos del mismo tamaño. El proceso de aprendizaje se realiza 5 veces en total. En cada caso, uno de los 5 conjuntos se utiliza como conjunto de validación y los otros 4 de

TABLA I
Resultados experimentales de clasificación individual y de validación cruzada.

	Clasificación Individual		**Validación cruzada**	
Conj. Entrenamiento	*Photo.net*	*CUHK*	*Photo.net*	*CUHK*
Conj. Validación	*Photo.net*	*CUHK*	*CUHK*	*Photo.net*
AvgStd	61,10%	60,27%	52,13%	51,79%
SizeFreq	63,02%	63,47%	53,33%	57,34%
AvgStd+ SizeFreq	66,05%	64,21%	54,22%	56,60%

entrenamiento. Por lo tanto, todos los patrones se usan una vez para la validación y 4 veces para el entrenamiento. Los resultados reportados en este experimento se refieren a los resultados de la validación.

En la columna "Clasificación Individual" de la Tabla I se muestran los resultados obtenidos. En ninguno de ellos se han alcanzado resultados comparables con los trabajos especializados en la búsqueda de características relacionadas con la calidad estética, lo cual resulta lógico ya que las características utilizadas son estimadores sencillos y de ámbito general.

En el segundo experimento se ha realizado una ligera modificación con respecto al primero. En lugar de usar la metodología descrita anteriormente se han utilizado los clasificadores entrenados en el experimento anterior para catalogar las imágenes de otro conjunto. De este modo comprobaremos si el clasificador entrenado con las imágenes pertenecientes a *Photo.net* clasifica correctamente las imágenes pertenecientes a *CUHK* y viceversa. Los resultados obtenidos en este nuevo experimento también se muestran en la Tabla I. Se puede observa como en el primer experimento se obtienen resultados cercanos al 61% en el peor de los casos, mientras que el segundo se obtiene un 57% en el mejor, con una diferencia media en torno al 10% usando las mismas características. Además, los resultados de la validación cruzada se acercan generalmente al 50% de acierto, lo cual parece indicar que el acierto conseguido podría deberse más al azar que a la capacidad del clasificador.

La diferencia al usar conjuntos diferentes de imágenes pertenecientes a fuentes similares en cuanto a concepto pero distintas en cuanto a contenido muestra una baja coherencia entre ambos conjuntos de imágenes. Estos resultados parecen concordar con nuestra sospecha inicial de que la utilización de los mismos conjuntos muestrales para entrenar y validar clasificadores estéticos de imágenes puede influir negativamente en su capacidad de generalización.

VIII. Discusión de los resultados

Este trabajo radica en comprobar si la elección de los datos muestrales es trivial para alcanzar cierto grado de generalización en tareas de clasificación estética. Hemos utilizado dos conjuntos de imágenes creados y categorizados en función a su calidad estética, la cual ha sido determinada por las valoraciones de los usuarios de distintos portales de fotografía. A partir de dichas imágenes hemos empleado distintos clasificadores binarios alimentados por valores generales que identifican a cada una de ellas. Los resultados obtenidos parecen indicar que ambos conjuntos empleados no se deben tomar como conjuntos representativos, aunque hayan sido utilizados en varios trabajos relacionados.

Sería interesante intentar buscar o crear previamente un conjunto de imágenes que se pueda adoptar como estándar general, aunque dicha tarea se presenta a nuestro entender como complicada. Por nuestra parte, en la actualidad estamos trabajando en la creación de un conjunto de imágenes propio obtenido exclusivamente de "DPChallenge.com". Dicho conjunto es diferente al de *CUHK*. Se compone de un total de 45,000 imágenes de diferentes temáticas que han participado en diversos concursos dentro del portal web. La media de las valoraciones de las imágenes es de 230 con una desviación estadística de 61, teniendo cada imagen al menos 20 votaciones de diferentes usuarios. Pretendemos utilizar dicho conjunto para continuar estudiando los problemas de generalización observados en este trabajo.

Agradecimientos

Los autores querrían agradecer a los siguientes entes públicos el haber aportado la financiación necesaria para llevar a cabo este trabajo: Portuguese Foundation for Science and Technology, proyecto de investigación PTDC/EIAEIA/115667/2009; Ministerio de Ciencia y Tecnología de España, proyecto de investigación TIN200806562/TIN; Xunta de Galicia, proyecto de investigación XUGAPGIDIT10TIC105008PR.

Referencias

[1] R. Datta, D. Joshi, J. Li, and J.Z. Wang, "Studying aesthetics in photographic images using a computational approach," *ECCV 2006. LNCS*, vol. 3953, pp. 288-301. Springer, Heidelberg, 2006.

[2] L. Wong, and K. Low, "Saliency-enhanced image aesthetics class prediction," *ICIP 2009*, pp. 997-1000. Los Alamitos, 2009.

[3] Y. Ke, X. Tang, and F. Jing, "The Design of High-Level Features for Photo Quality Assessment," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 419-426, 2006.

[4] Y. Luo, and X. Tang, "Photo and video quality evaluation: Focusing on the subject," *ECCV 2008, Part III. LNCS*, vol. 5304, pp. 386-399. Springer, Heidelberg, 2008.

[5] J. Romero, P. Machado, A. Carballal, and O. Osorio, "Aesthetic Classification and Sorting Based on Image Compression," *Lecture Notes in Computer Science, Applications of Evolutionary Computation*, vol. 6625, pp. 394-403. Springer, Heidelberg, 2011.

[6] R. Datta, J. Li and, J. Z. Wang, "Algorithmic inferencing of aesthetics and emotion in natural images: An exposition," *ICIP 2008*, pp. 105-108, 2008.

[7] J. Romero, P. Machado, A. Carballal, A. Santos, "Using complexity estimates in aesthetic image classification",

Journal of Mathematics and the Arts, vol. 6, 2-3, 2012.
[8] R. Arnheim, *Art and Visual Perception, a psychology of the creative eye*, London: Faber and Faber, 1956.
[9] P. Machado, and , A. Cardoso, "Computing aesthetics," *SBIA 1998. LNCS (LNAI)*, vol. 1515, pp. 219-229. Springer, Heidelberg, 1998.
[10] P. Machado, J. Romero, and B. Manaris, "Experiments in Computational Aesthetics," *The Art of Artificial Evolution*. Springer, Heidelberg, 2007.
[11] http://www.cs.waikato.ac.nz/ml/weka/
[12] http://www.csie.ntu.edu.tw/~cjlin/libsvm/